Big Data Security and Privacy

Big Data Management and Analytics

Overview
Lecture #1-7

Dr. Latifur Khan
Dr. Bhavani Thuraisingham

Introduction to Big Data and Big Data Management

What is Big Data?

What makes data, "Big" Data?

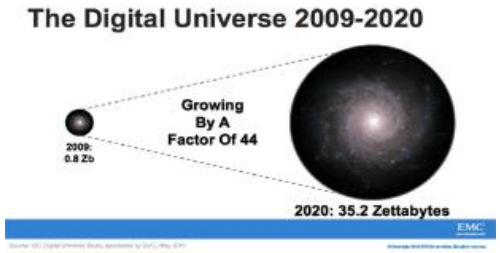
Big Data Definition

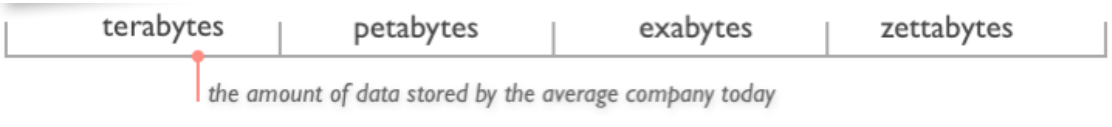
- No single standard definition…

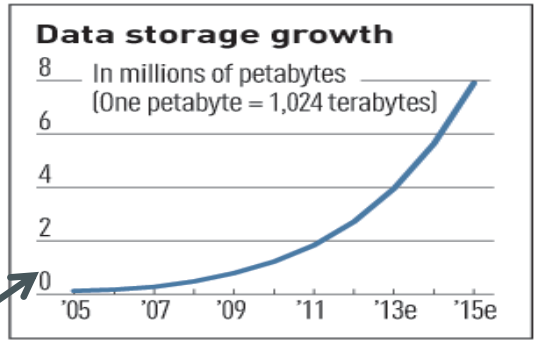
"***Big Data***" is data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it…

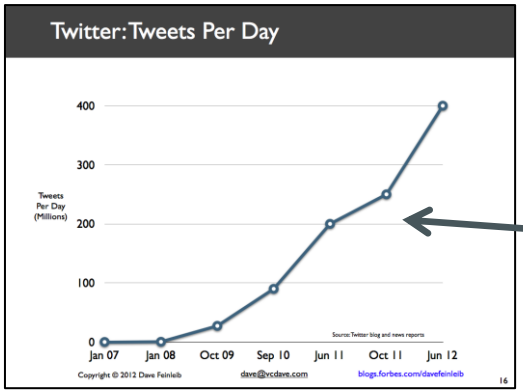
Characteristics of Big Data: 1-Scale (Volume)

- **Data Volume**
  - 44x increase from 2009 2020
  - From 0.8 zettabytes to 35zb
- Data volume is increasing exponentially





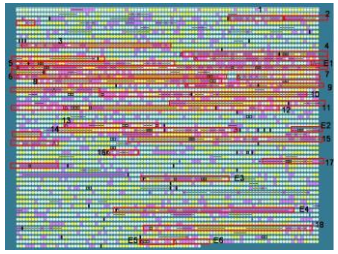


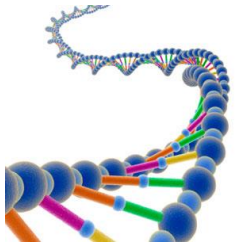


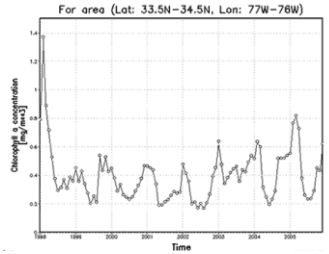
Exponential increase in collected/generated data

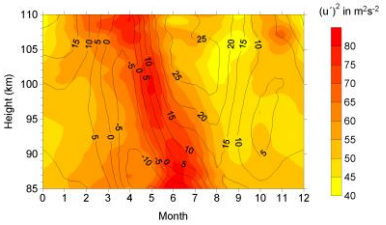
Characteristics of Big Data: 2-Complexity (Variety)

- Various formats, types, and structures
- Text, numerical, images, audio, video, sequences, time series, social media data, multi-dim arrays, etc…
- Static data vs. streaming data
- A single application can be generating/collecting many types of data









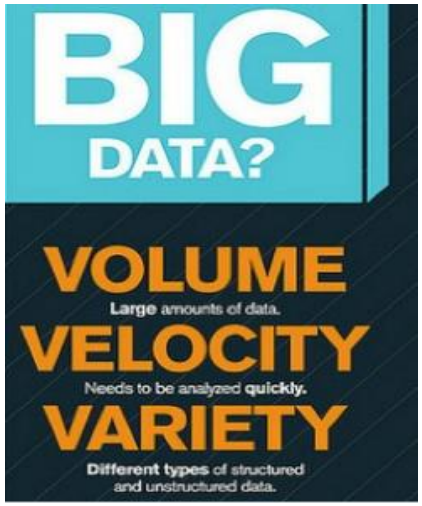
To extract knowledge➔ all these types of data need to linked together

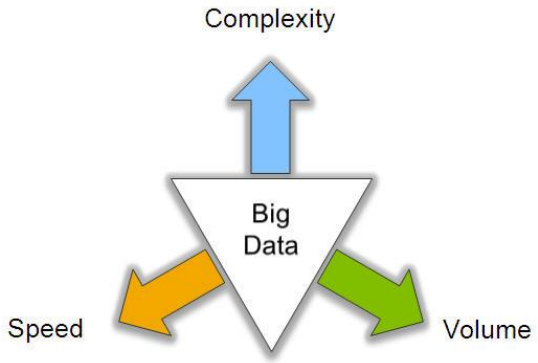
Characteristics of Big Data: 3-Speed (Velocity)

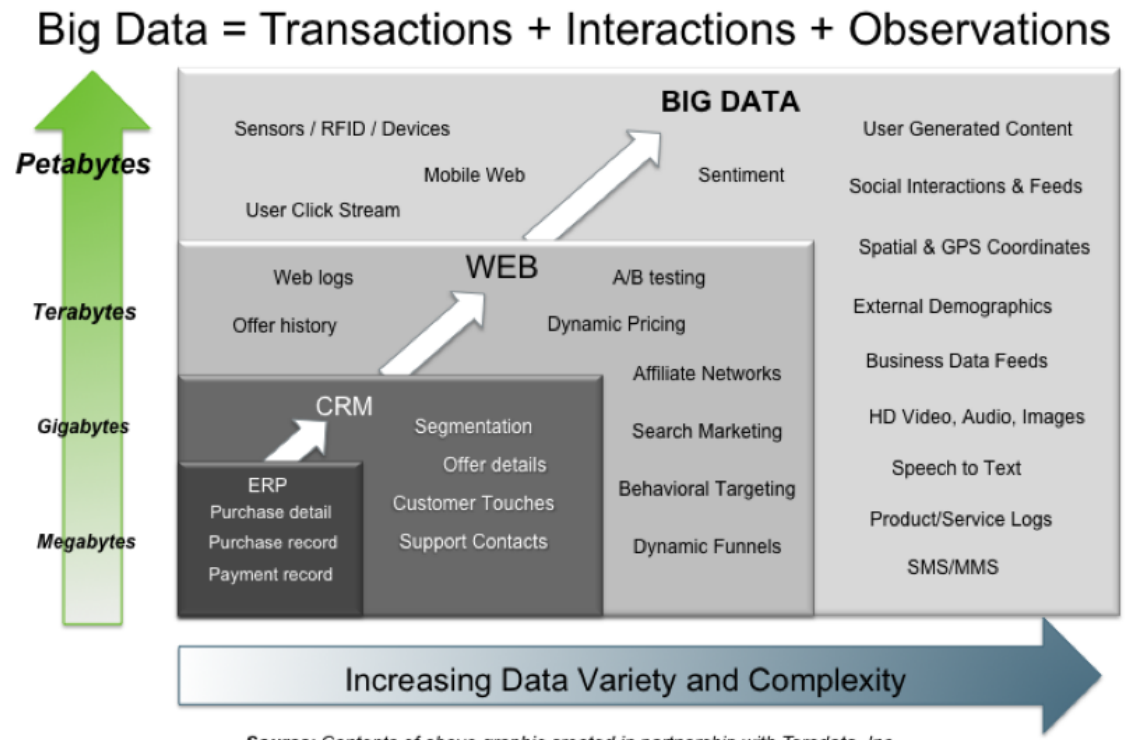
- Data is begin generated fast and need to be processed fast
- Online Data Analytics
- Late decisions ➔ missing opportunities

- **Examples**
  - **E-Promotions:** Based on your current location, your purchase history, what you like ➔ send promotions right now for store next to you
  - **Healthcare monitoring:** sensors monitoring your activities and body ➔ any abnormal measurements require immediate reaction

Big Data: 3V's

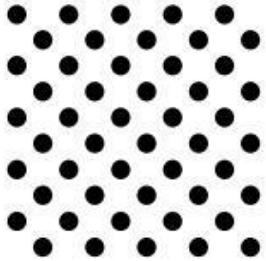
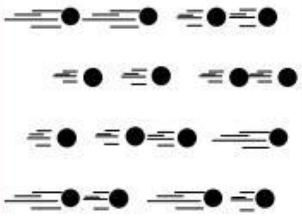
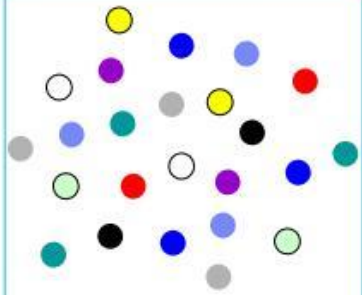
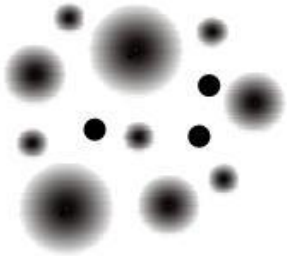




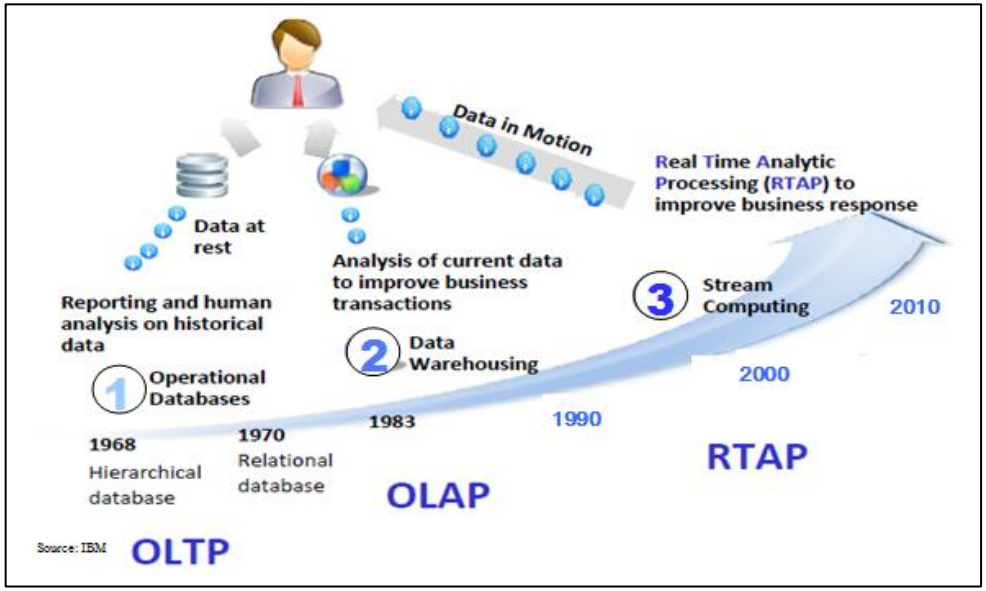


***Source**: Contents of above graphic created in partnership with Teradata, Inc.*

Some Make it 4V's

Volume	Velocity	Variety	Veracity*
			
Data at Rest	**Data in Motion**	**Data in Many Forms**	**Data in Doubt**
Terabytes to exabytes of existing data to process	Streaming data, milliseconds to seconds to respond	Structured, unstructured, text, multimedia	Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

Harnessing Big Data



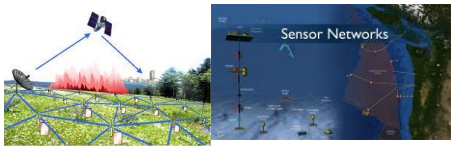
- **OLTP:** Online Transaction Processing (DBMSs)
- **OLAP:** Online Analytical Processing (Data Warehousing)
- **RTAP:** Real-Time Analytics Processing (Big Data Architecture & technology)

Who's Generating Big Data

Social media and networks
(all of us are generating data)

Scientific instruments
(collecting all sorts of data)

Mobile devices
(tracking all objects all the time)



Sensor technology and networks
(measuring all kinds of data)

- The progress and innovation is no longer hindered by the ability to collect data
- But, by the ability to manage, analyze, summarize, visualize, and discover knowledge from the collected data in a timely manner and in a scalable fashion

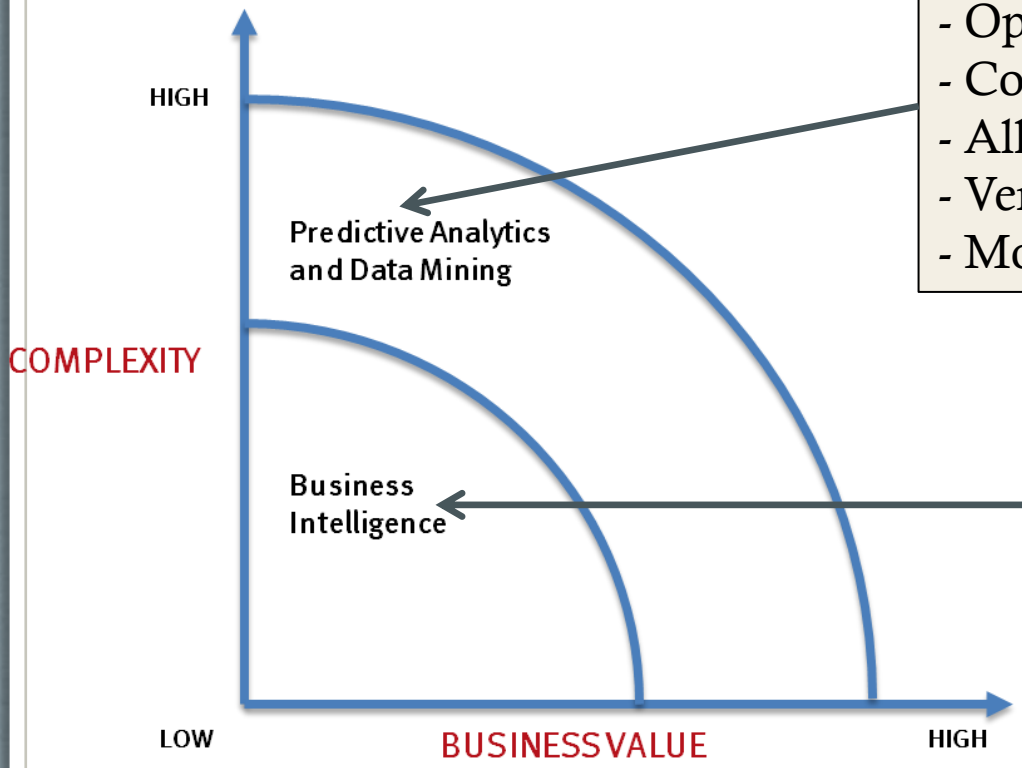
The Model Has Changed…

- **The Model of Generating/Consuming Data has Changed**

Old Model: Few companies are generating data, all others are consuming data

New Model: all of us are generating data, and all of us are consuming data

What's driving Big Data

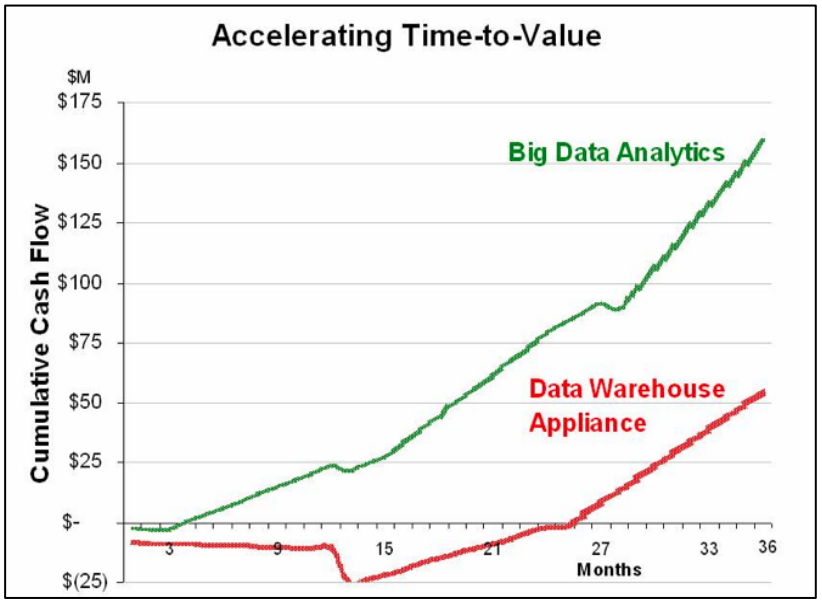


- Optimizations and predictive analytics
- Complex statistical analysis
- All types of data, and many sources
- Very large datasets
- More of a real-time

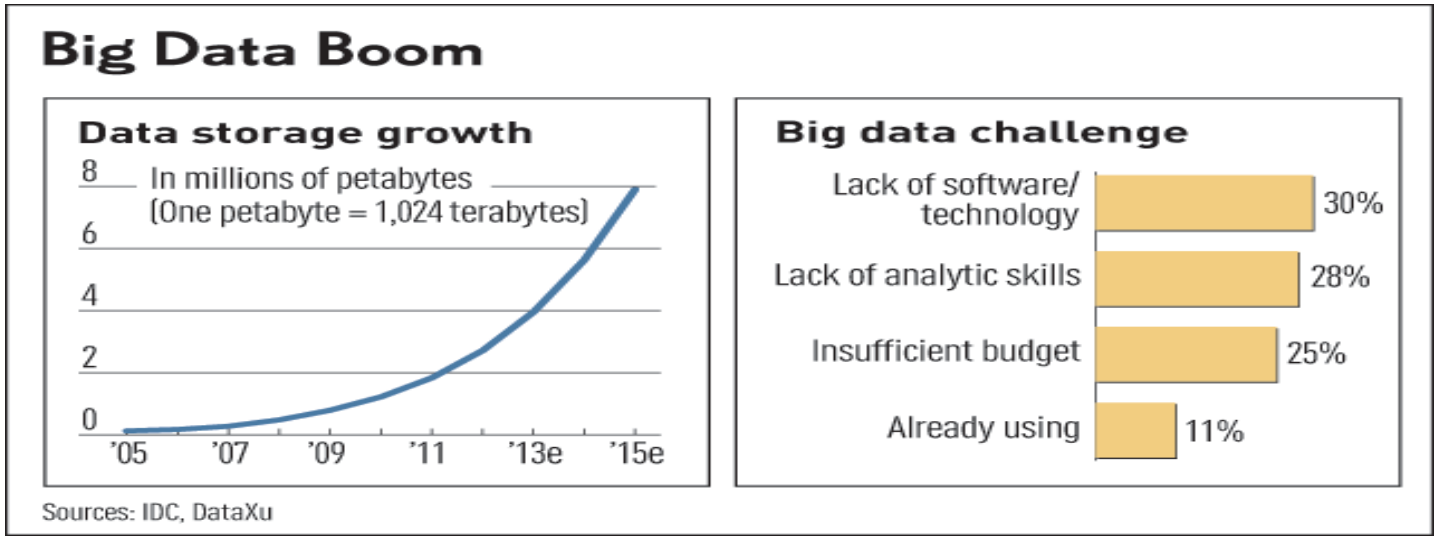
- Ad-hoc querying and reporting
- Data mining techniques
- Structured data, typical sources
- Small to mid-size datasets

Value of Big Data Analytics

- Big data is more real-time in nature than traditional DW applications
- Traditional DW architectures (e.g. Exadata, Teradata) are not well-suited for big data apps
- Shared nothing, massively parallel processing, scale out architectures are well-suited for big data apps



Challenges in Handling Big Data



- **The Bottleneck is in technology**
  - New architecture, algorithms, techniques are needed
- **Also in technical skills**
  - Experts in using the new technology and dealing with big data

What Technology Do We Have For Managing Big Data ??

Big Data Landscape

Vertical Apps

PREDICTIVE POLICING
bloomreach
MYRRIX

Log Data Apps

splunk> loggly sumologic

Ad/Media Apps

rocketfuel
collective[i]
bluefin
Recorded Future
LuckySort
Media Science
TURN
DataXu

Business Intelligence

ORACLE | Hyperion
SAP Business Objects RJMetrics
Microsoft | Business Intelligence
IBM COGNOS birst
MicroStrategy
Autonomy
bime
QlikView
Chart.io
GoodData

Analytics and Visualization

tableau
Palantir
OPERA
metaLayer
METAMARKETS
dataspora
centrifuge
TERADATA ASTER
SAS TIBCO
KARMASPHERE
panopticon
Real-Time Visual Data Analysis
pentaho
Datameer
platfora
ClearStory
CIRRO
alteryx
visual.ly
AYATA

Data As A Service

factual
kaggle
knoema beta
GNIP
DATASIFT
Windows Azure Marketplace
INRIX
LexisNexis
LOQATE

Analytics Infrastructure

Hortonworks
VERTICA
MAPR
cloudera
INFOBRIGHT
PARACCEL
EMC²
GREENPLUM
NETEZZA
kognitio
DATASTAX
EXASOL
calpont

Operational Infrastructure

COUCHBASE
10gen
TERADATA
HADAPT
TERRACOTTA
VoltDB
MarkLogic
INFORMATICA

Infrastructure As A Service

amazon web services
Windows Azure
infochimps
Google BigQuery

Structured Databases

ORACLE
MySQL
Microsoft SQL Server
PostgreSQL
IBM DB2
SYBASE
memsql

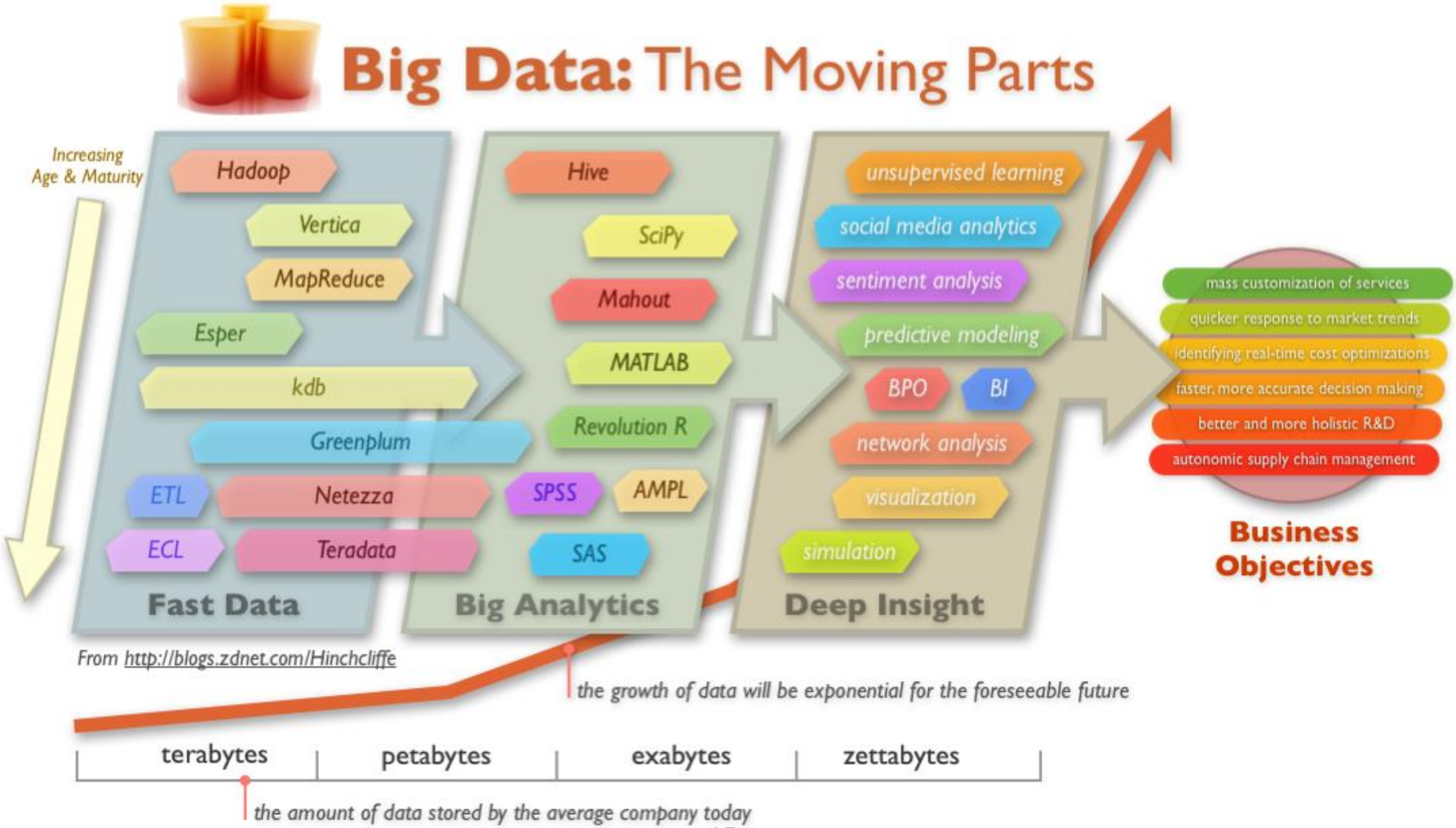
Technologies

hadoop
hadoop MapReduce
mahout
APACHE HBASE
Cassandra

dave@vcdave.com

blogs.forbes.com/davefeinleib

Big Data Technology



Big Data Analytics

Big Data Analytics Course: Prof. Latifur Khan

- We focus on ***Hadoop/MapReduce technology, NoSQL (key-Value), BigTable, Cassandra, SPARK, Stream…***
- **Learn the platform (how it is designed and works)**
  - How big data are managed in a scalable, efficient way
- **Learn writing Hadoop jobs/SPARK in different languages**
  - Programming Languages: Java, C, Python, Scala
  - High-Level Languages: Apache Pig, Hive, CQL
- **Learn advanced analytics tools on top of Hadoop**
  - Mahout: Data mining and machine learning tools over big data
  - Applications: Recommendations
- Graph Processing: Pregel (Giraf)

Recommendation Systems

Search

Recommendations

Items

Products, web sites, blogs, news items, …

Examples:

amazon.com

PANDORA

StumbleUpon

del.icio.us

NETFLIX

movielens
helping you find the *right* movies

last·fm
the social music revolution

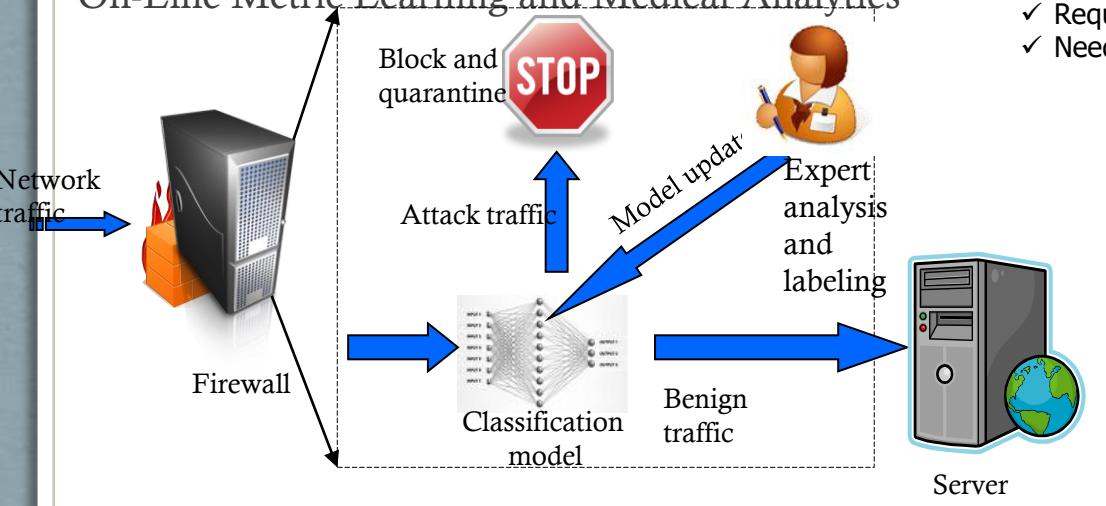
Google News

YouTube

XBOX LIVE

Big Data Stream Analytics and Its Applications:

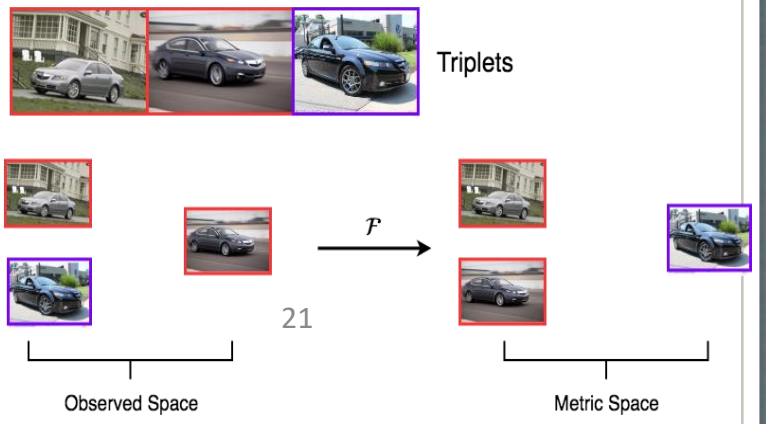
- Data Streams: Single Stream and Multi-stream
- Novel Class Detection & Zero Day Attack
- Transfer Learning & Domain Adaptation
- Big Text Analytics: Political Report Analysis***
- On-Line Metric Learning and Medical Analytics***



Amazon Reviews

✓ Require model update
✓ Need truth labels

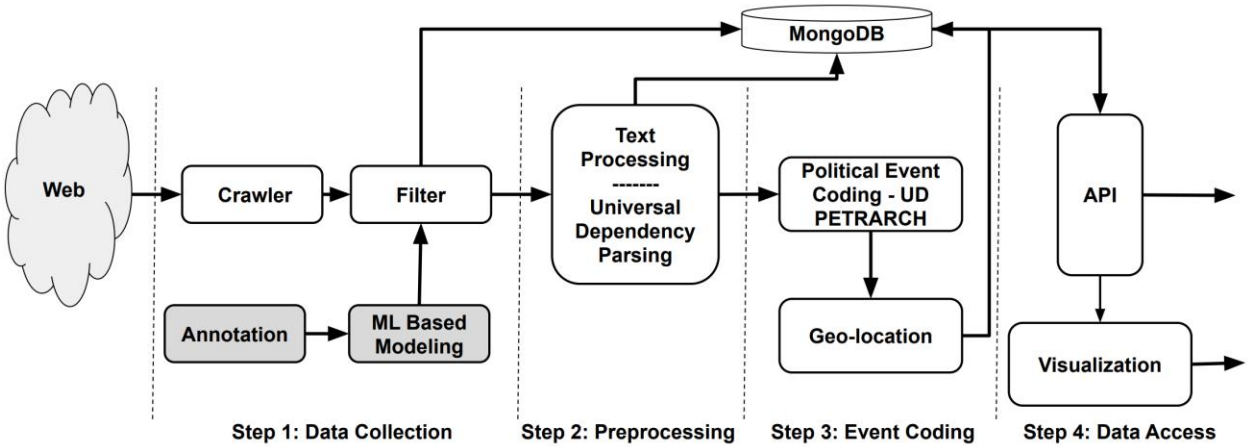
✓ Leverage available truth label
✓ Predict on target data



*** UG can join

Big Data Text Analytics

- **Data Science for Political Sciences (Project for Students)**
- **Sustainable Infrastructure development for Automatic Event Coding in real-time**
  - **Ontology extension for Automatic Event Coder**
  - **Multilingual Support with ontology translation**
  - **Data Sharing and Visualization**



Next Steps:

Big Data Security and Privacy

Integrates

Data Security and Privacy with

Big Data Management and Analytics

Big Data Security and Privacy Course: Prof. Bhavani Thuraisingham

- **Massive amounts of data being collected and analyzed**
- **Big Data has to be Secured**
- **It is possible to infer private information**
- **Security and Privacy Major Consideration for Big Data**
- **Big Data Techniques can also be attacked**
- **Big Data Security and Privacy aka Secure Data Science**

NSF Workshop on Big Data Security and Privacy

Report Summary

Bhavani Thuraisingham
The University of Texas at Dallas (UTD)

February 19, 2015

Acknowledgement

- NSF SaTC Program for support
  - Chris Clifton and Jeremy Epstein
- Workshop Working Group Chairs
  - Elisa Bertino, Purdue University
  - Murat Kantarcioglu, University of Texas at Dallas
- Workshop Keynote Speakers, Presenters and Participants
- Workshop Coordinator
  - Rhonda Walls, University of Texas at Dallas
- Workshop Report Link
  - http://csi.utdallas.edu/events/NSF/NSF-workhop-Big-Data-SP-FINAL.pdf

Organization of the Workshop

- Workshop held at the University of Texas at Dallas, September 16 and 17, 2014
- Invited keynote speakers and participants from interdisciplinary research communities
  - Computer and Data Scientists, Cyber Security Researchers, Social Scientists, Application Specialists from Natural Sciences
  - Academia, Industry and Government
- Two Major Objectives
  - Security and Privacy Issues for Big Data Applications
  - Big Data Management and Analytics Techniques for Cyber Security Applications
- Keynote Presentations, Position Paper Presentation and Two Break out sessions to address the major objectives; Breakout session discussion reports
- Submitted to the National Privacy Research Strategy, October 16, 2014
- Workshop Report published on February 9, 2015

Big Data Management and Analytics: Pros and Cons

- Due to technological advances and novel applications it is possible to capture, process, analyze large amounts of data for security tasks
- Such tasks include user authentication, access control, anomaly detection, user monitoring, and protection from insider threat
- By analyzing and integrating data collected on the Web, one can identify connections and relationships among individuals that may in turn help with homeland protection, disease outbreaks, ….
- Collected data, even if anonymized by removing identifiers, when linked with other data, may lead to re-identifying the individuals
- Security tasks such as authentication and access control may require detailed information about users (e.g., multi-factor authentication)
- This information if misused or stolen can lead to privacy breaches.

Examples of Privacy Enhancing Techniques

- Numerous privacy-enhancing techniques have been proposed ranging from cryptographic techniques to data anonymization

- Such techniques either do not scale for large datasets and/or do not address the problem of reconciling security with privacy.

- Few approaches focus on efficiently reconciling security with privacy; these can be grouped as follows:

  Privacy-preserving data/record matching
  Privacy-preserving collaborative data mining
  Privacy-preserving biometric authentication

Privacy Preserving Record Matching

- Record matching is performed across different data sources with the aim of identifying shared common information
- Matching records from different data sources may conflict with privacy requirements of the individual data sources.
- Recent approaches include data transformation and mapping into vector spaces, and combination of secure multiparty computation and data sanitization (e.g., differential privacy and k-anonymity)
- Need privacy-preserving techniques suitable for complex matching problems (e.g., semantic matching).
- Need security models and definitions supporting security analysis and proofs for solutions combining different security techniques

Privacy Preserving Collaborative Data Mining

- Data mining is typically performed on centralized data warehouses
- Centrally collecting data poses privacy and confidentiality concerns multi-organizational data
- Distributed collaborative approaches where organizations retain their own datasets and cooperate to learn the global data mining results
- These techniques are still inefficient
- Need novel approaches based on cloud computing and new cryptographic primitives

Privacy Preserving Biometrics Authentication

Record biometrics templates of enrolled users and match with the templates provided by users during authentication

Templates of user biometrics represent sensitive information

Recent approaches address the problem by using a combination of perceptual hashing techniques, classification techniques, and zero-knowledge proof of knowledge (ZKPK) protocols

Need research to reduce the false rejection rates

Need approaches for authentication and based on the recent homomorphic encryption techniques

Multi-Objective Optimization Framework for Data Privacy

- Attempts at coming up with a privacy solution/definition that can address different scenarios; but there is no one size fits all solution for data privacy.

- Data Utility needs to be included

  For privacy-preserving classification models, 0/1 loss could be a good utility measure.

  For privacy-preserving record linkage, F1 score could be a better choice.

- Need to understand the right definitions of privacy risk

  For data sharing, probability of re-identification given certain background knowledge may be a right measure of privacy risk.

  $\varepsilon=1$ may be an appropriate risk for differentially private data mining models.

- The computational, storage and communication costs of the protocols need to be considered.

Multi-Objective Optimization Framework for Data Privacy

- Need to develop a multi-objective framework where different dimensions could be emphasized:

- **Maximize utility, given risk and costs constraints:** This would be suited for scenarios where limiting certain privacy risks are paramount.

- **Minimize privacy risks, given the utility and cost constraints**: In some scenarios significant degradation of the utility may not be allowed.

  The parameter values of the protocol (e.g., ε in differential privacy) are chosen in such way to maximize privacy given utility constraints.

- **Minimize cost, given the utility and risk constraints:** May need to find the protocol parameter settings that may allow for the least expensive protocol that can satisfy the utility and risk constraints.

Research Challenges and Multidisciplinary approaches

- Data Confidentiality: For access control systems for Big Data we need approaches for:
  - *Merging large number of access control policies*. Policies (e.g., sticky policies) need to be integrated and conflicts solved.
  - *Automatically administering authorizations for big data and in particular for granting permissions*. Need techniques to automatically grant authorization, possibly based on the user digital identity, profile, and context, data contents and metadata.
  - *Enforcing access control policies on heterogeneous multi-media data*. Supporting content-based access control requires understanding the contents of protected data
  - *Enforcing access control policies in big data stores*. Need to efficiently inject access control policies into submitted jobs (e.g., MapReduce)
  - *Automatically designing, evolving, and managing access control policies*. When dealing with dynamic environments there is a need to automatically design and evolve policies

Research Challenges and Multidisciplinary approaches

- Privacy-preserving data correlation techniques Relevant issues that need to be investigated include:
  - *Support for both personal privacy and population privacy*.

    Need to understand what is extracted from the data as this may lead to discrimination.

    When dealing with security with privacy, it is important to understand the tradeoff of personal privacy and collective security.
  - *Efficient and scalable privacy-enhancing techniques*.

    Need to parallelize the privacy enhancing techniques
  - *Usability of data privacy policies*.

    Policies must be easily understood by users.
  - *Approaches for data services monetization*.

    Instead of selling data, organizations owning datasets can sell privacy-preserving data analytics services based on these datasets.

Research Challenges and Multidisciplinary approaches

- *Data publication*.

  Perhaps abandon the idea of publishing data and use data in a controlled environment

- *Privacy implication on data quality*.

  Studies have shown that people lie especially in social networks
  This results in a decrease in data quality that affects decisions

- *Risk models.*

  (a) Big data can increase privacy risks; (b) Big data can reduce risks in many domains (e.g. national security).
  Need models for these two types of risks

- *Data ownership.*

  Perhaps replace “who is the owner of the data” with the concept of stakeholder(s).
  Each stakeholder may have different possibly conflicting objectives and this can be modeled according to multi-objective optimization.

Research Challenges and Multidisciplinary approaches

- *Human factors*.
  All solutions proposed for privacy and for security with privacy need to consider human involvement

- *Data lifecycle framework.*

  A comprehensive approach to privacy for big data needs to be based on a systematic data lifecycle approach.

  Relevant phases include:

  Data acquisition – Need mechanisms and tools to prevent devices from acquiring data about other individuals

  Data sharing – users need to be informed about data sharing/transfers to other parties.

Addressing the above challenges requires multidisciplinary research drawing from many different areas, including computer science and engineering, information systems, data science and statistics, risk models, economics, social sciences, political sciences, public policy, human factors, psychology, law.

Big Data Management and Analytics for Security

What is different about Big Data analytics (BDA) for Cyber security?

BDA for cyber security needs to deal with adaptive, malicious adversary

BDA for cyber security needs to operate in high volume and high noise environments

Need BDA tools to integrate data from hosts, networks, social media, bug reports, mobile devices, and internet of things to detect attacks.

What is the right BDA architecture for Cyber Security?:

Do we need different types of BDA system architectures for cyber security?

Should existing BDA system architectures be adapted for cyber security needs?

Need real time data analysis

Data Sharing for BDA for Cyber Security:

Cyber security data needs to be shared both inside the organization and among organizations

Need common languages and infrastructure to capture and share such cyber security data.

.

Big Data Management and Analytics for Security

- *BDA for Preventing Cyber Attacks:*

  BDA systems that capture provenance information can potentially detect attacks before too much sensitive information is disclosed.

  Need to build provenance-aware BDA systems

  BDA tools for cyber security can potentially mine useful attacker information (motivations, capabilities, modus operandi)

- *BDA for Digital Forensics:*

  BDA techniques could be used for digital forensics by combining or linking different data sources.

  The main challenge is identifying the right data sources for digital forensics.

- *BDA for Understanding the Users of the Cyber Systems:*

  BDA could be used to mine human behavior to learn how to improve the systems.

  BDA techniques could be used to understand and build normal behavior models per user to find significant deviations

Summary and Directions

As massive amounts of data are being collected, stored, manipulated, merged, analyzed, and expunged, security and privacy concerns will explode.

Investigated the issues surrounding Big Data Security and Privacy as well as applying Big Data Management and Analytics for Cyber Security.

Need scalable and practical solutions

Need to develop technologies guided by policies to address security and privacy issues throughout the lifecycle of the data.

Need to understand not only the societal impact of data collection, use and analysis, also need to formulate appropriate laws and policies for such activities.

Need Intra and Inter-Agency Programs for Big Data Security and Privacy

Contact

- **Ms. Rhonda Walls**
  rhonda.walls@utdallas.edu, (972) 883-2731
- **Dr. Bhavani Thuraisingham**
  bhavani.thuraisingham@utdallas.edu, (972) 883-4738
- **Follow us @CyberUTD**