Introduction to Big Data

Bal Krishna Nyaupane

Assistant Professor

Department of Electronics and Computer Engineering

Institute of Engineering, Tribhuvan University

bkn@wrc.edu.np

Chapter 1: Introduction to Big Data

- Big Data Overview
- Background of Data Analytics
- Role of Distributed System in Big Data
- Role of data Scientist
- Current Trend in Big Data Analytics

1 byte (B)	8 bit
1 kilobyte (K/Kb)	2^{10} byte = 1024 byte
1 megabyte (M/Mb)	2^{20} byte = 1024 Kb
1 gigabyte (G/Gb)	2^{30} byte = 1024 Mb
1 terabyte (T/Tb)	2^{40} byte = 1024 Gb
1 petabyte (P/Pb)	2^{50} byte = 1024 Tb
1 exabyte (E/Eb)	2^{60} byte = 1024 Pb
1 zettabyte (Z/Zb)	2^{70} byte = 1024 Eb
1 yottabyte (Y/Yb)	2^{80} byte = 1024 Zb

Dawn of the Big Data Era

- Over the past 20 years, data has increased in a large scale in various fields.
- According to a report from ***International Data Corporation (IDC)*, in 2011**, the overall created and copied data volume in the world **was 1.8ZB**, which has increased by nearly ***nine times within 5 years***. Such figure will ***double at least every other 2 years*** in the near future.
- Big data are often covered in public media, including ***The Economist*** , ***New York Times, and National Public Radio***. *Two premier scientific journals,* ***Nature and Science****, also started special columns to discuss the importance and challenges of* big data.
- Big data also brings new ***opportunities for discovering new values***, helps us to gain an ***in-depth understanding of the hidden values***, and incurs new challenges.

Dawn of the Big Data Era

- Recently, the rapid growth of big data mainly comes from people's daily life, especially related to the service of Internet companies.
- For Example,
  - ***Facebook generates log data of over 10 Petabyte (PB) per month***
  - Baidu, a Chinese company, processes data of tens of PB.
  - Taobao, a subsidiary of Alibaba, generates data of tens of Terabyte (TB) on online trading per day.
  - The rapid growth of ***cloud computing and the Internet of Things (IoT)*** further promote the sharp growth of data.
  - In the paradigm of IoT, ***sensors all over the world are collecting and transmitting data*** which will be stored and processed in the cloud.

Who's Generating Big 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

Mobile devices

(tracking all objects all the time)

Social media and networks

(all of us are generating data)

Scientific instruments

(collecting all sorts of data)

Sensor technology and networks

(measuring all kinds of data)

CERN's Large Hydron Collider (LHC) generates 15 PB a year (tunnel for measurement of protons , neutrons collision at the speed of light) developed for study purpose..

- DATA Today…


30 billion RFID tags today (1.3B in 2005)
4.6 billion camera phones world wide
100s of millions of GPS enabled devices sold annually
2+ billion people on the Web by end 2011
? TBs of data every day
25+ TBs of log data every day
76 million smart meters in 2009… 200M by 2014
http://www.
Google
ooTalk
Google Analytics
YouTube

The Phenomenon of Big Data

1.8ZB

Data generated during 2 days in 2011
(larger than the accumulated amount of data generated from the origin of civilization to 2003)

750 million

The amount of pictures uploaded to Facebook

966PB

In 2009, the storage capcity of American manufacturing industry

209 billion

The number of RFID tags in 2021
(12 million in 2011)

200+TB

Data downloaded during a computer geek's 2450 thousand hours

200PB

The amount of data generated by a smart urban project in China

800 billion dollars

Personal location data in 10 years

300 billion dollars

Medical expense saving by big data analysis in America

$32+B

The purchase amount of the 4 big companies since 2010

"Data are becoming the new raw material of business: Economic input is almost equivalent to capital and labor" -<<*Economist*>>, 2010

"Information will be 'the 21th Century oil." - *Gartner company, 2010*

How much data?

- 500 Million Tweets sent each day!
- More than 4 Million Hours of content uploaded to Youtube every day!
- 3.6 Billion Instagram Likes each day.
- 4.3 BILLION Facebook messages posted daily!
- 5.75 BILLION Facebook likes every day.
- 40 Million Tweets shared each day!
- 6 BILLION daily Google Searches!

https://blog.microfocus.com/how-much-data-is-created-on-the-internet-each-day/

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

Introduction to Big Data

- **Big data** is data sets that are so **big and complex** that traditional data-processing application software are **inadequate to deal with them**.
- **Big data is** a term applied to data sets whose **size or type is beyond the ability of traditional relational databases** to capture, manage, and process the data **with low-latency**.
- **Big data analytics** is the use of advanced analytic techniques against **very large, diverse data sets** that include **structured, semi-structured and unstructured data**, from different sources, and in different **sizes from Terabytes to Zettabytes.**

Structured Data

- This is the data which ***is an organized form(i.e. in rows and columns)*** and can be easily used by computer program. ***Relationships exist between entities of data, such as classes and their objects.***
- Think of data that fits neatly within fixed fields and columns in relational databases and spreadsheets.
- ***Structured data is highly organized and easily understood by machine language***.
- Those working within relational databases can input, search, and manipulate structured data relatively quickly.
- ***Examples of structured data include*** names, dates, addresses, credit card numbers, stock information, geo-location, and more.
- Data stored in databases is an example of structured data.

Unstructured Data

- Unstructured data is information that ***either does not have a predefined data model or is not organized in a pre-defined manner.***
- Unstructured information ***is typically text-heavy***, but may ***contain data such as dates, numbers, and facts as well.***
- ***Common examples of unstructured data include*** text, video, audio, mobile activity, social media activity, satellite imagery, surveillance imagery, body of emails – the list goes on and on.
- Unstructured data is ***difficult to deconstruct because it has no pre-defined model***, meaning it cannot be organized in relational databases.
- Instead, non-relational, or NoSQL databases, are best fit for managing unstructured data.

Semi structured Data

- Semi-structured data is ***a form of structured data that does not conform with the formal structure of data models associated with relational databases*** or other forms of data tables, ***but nonetheless contain tags or other markers to separate semantic elements*** and enforce hierarchies of records and fields within the data.
- Therefore, it is also ***known as self-describing structure***.
- Examples of semi-structured data include ***JSON and XML are forms of semi-structured data.***
- The reason that this third category exists (between structured and unstructured data) is because semi-structured data is considerably easier to analyze than unstructured data.
- Many Big Data solutions and tools have the ability to 'read' and process either JSON or XML.

Big Data Characteristics

- As a matter of fact, big data has been defined as early as 2001. Doug Laney, an analyst of META defined challenges and opportunities brought about by the increased data with a ***3Vs model***, ***i.e., the increase of Volume, Velocity, and Variety, in a research report.***

- Although such a model was not originally used to define big data, Gartner and many other enterprises, including IBM and some research departments of Microsoft still used the "***3Vs" model to describe big data within the following 10 years.***

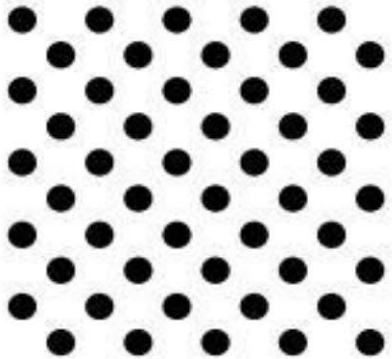
3Vs of Big Data

Big Data Technology is a **new set of approaches** for analyzing data sets that were not previously accessible because they posed challenges across one or more of the "3 V's" of Big Data

- **Volume -** too Big – Terabytes and more of Credit Card Transactions, Web Usage data, System logs
- **Variety -** too Complex – truly unstructured data such as Social Media, Customer Reviews, Call Center Records
- **Velocity -** too Fast - Sensor data, live web traffic, Mobile Phone usage, GPS Data

Some Make it 4V's

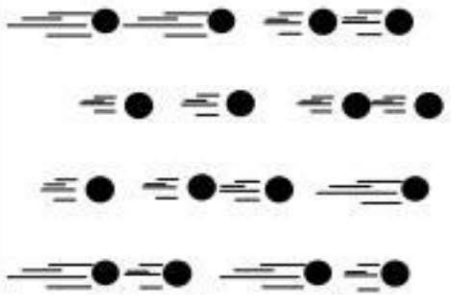
Volume



Data at Rest

Terabytes to exabytes of existing data to process

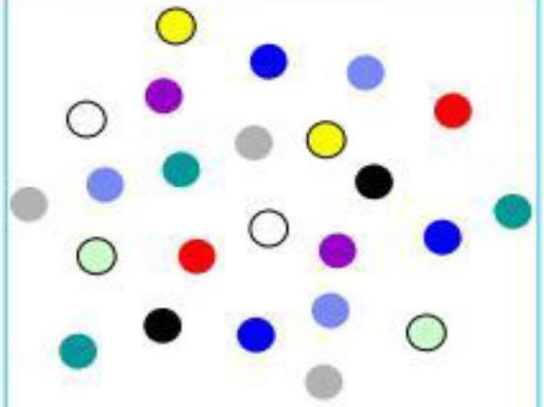
Velocity



Data in Motion

Streaming data, milliseconds to seconds to respond

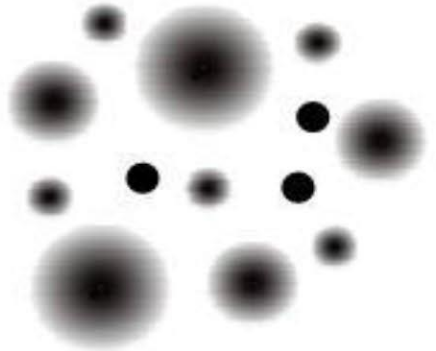
Variety



Data in Many Forms

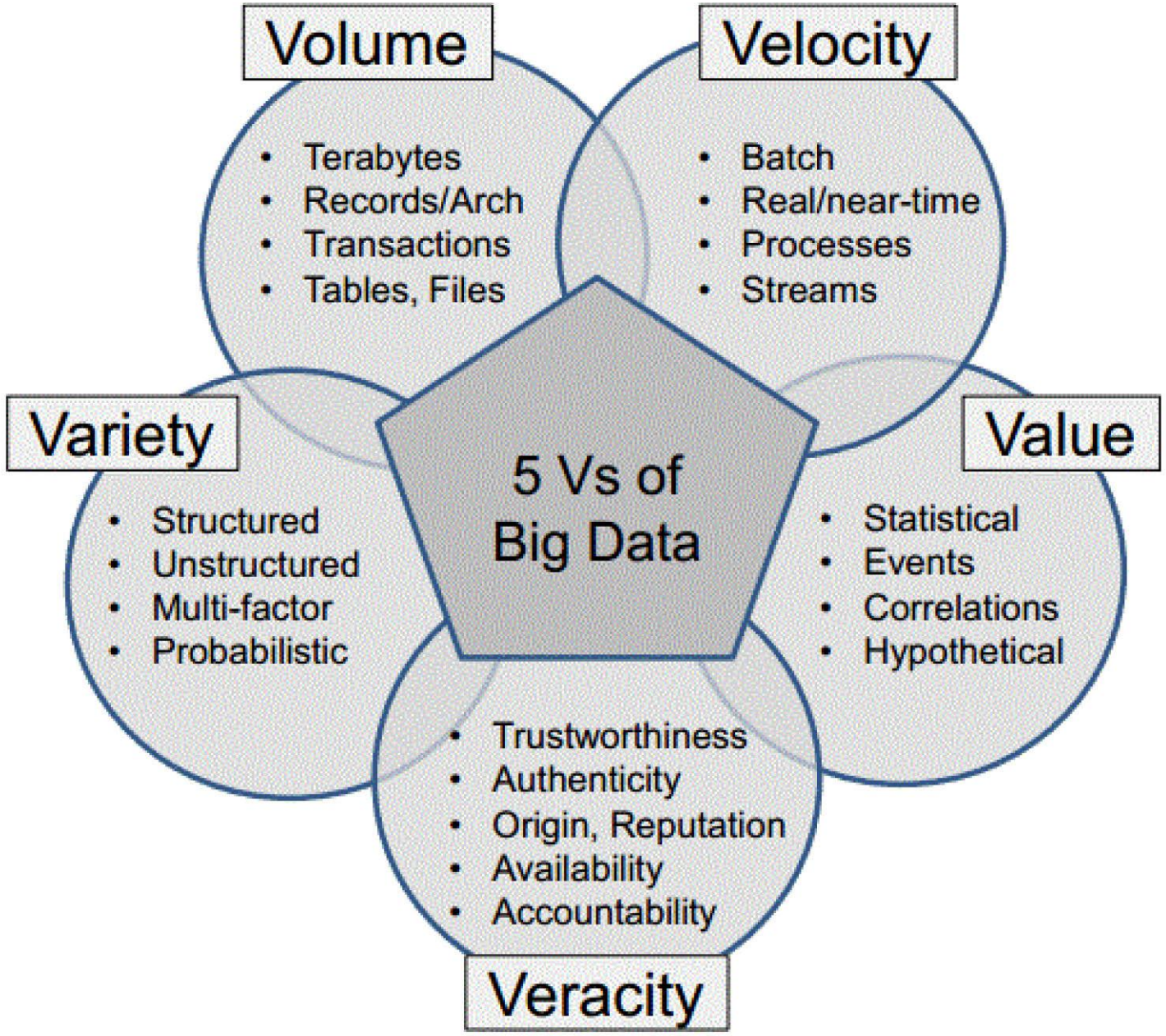
Structured, unstructured, text, multimedia

Veracity*



Data in Doubt

Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations


Volume
Terabytes
Records/Arch
Transactions
Tables, Files
Velocity
Batch
Real/near-time
Processes
Streams
Variety
Structured
Unstructured
Multi-factor
Probabilistic
5 Vs of
Big Data
Value
Statistical
Events
Correlations
Hypothetical
Trustworthiness
Authenticity
Origin, Reputation
Availability
Accountability
Veracity

Volume

- Refers to the vast amounts of data generated every second.
- We are not talking Terabytes but Zettabytes or Brontobytes.
- If we take all the data generated in the world between the beginning of time and 2008, the same amount of data will soon be generated every minute. This makes most data sets too large to store and analyze using traditional database technology.
- New big data tools use distributed systems so that we can store and analyze data across databases that are dotted around anywhere in the world.

Variety

- Refers to the different types of data we can now use.
- In the past we only focused on structured data that neatly fitted into tables or relational databases, such as financial data.
- In fact, 80% of the world's data is unstructured (text, images, video, voice, etc.)
- With big data technology we can now analyze and bring together data of different types such as messages, social media conversations, photos, sensor data, video or voice recordings.

Velocity

Refers to the speed at which new data is generated and the speed at which data moves around. Just think of social media messages going viral in seconds.

Technology allows us now to analyze the data while it is being generated (sometimes referred to as in-memory analytics), without ever putting it into databases.

Veracity

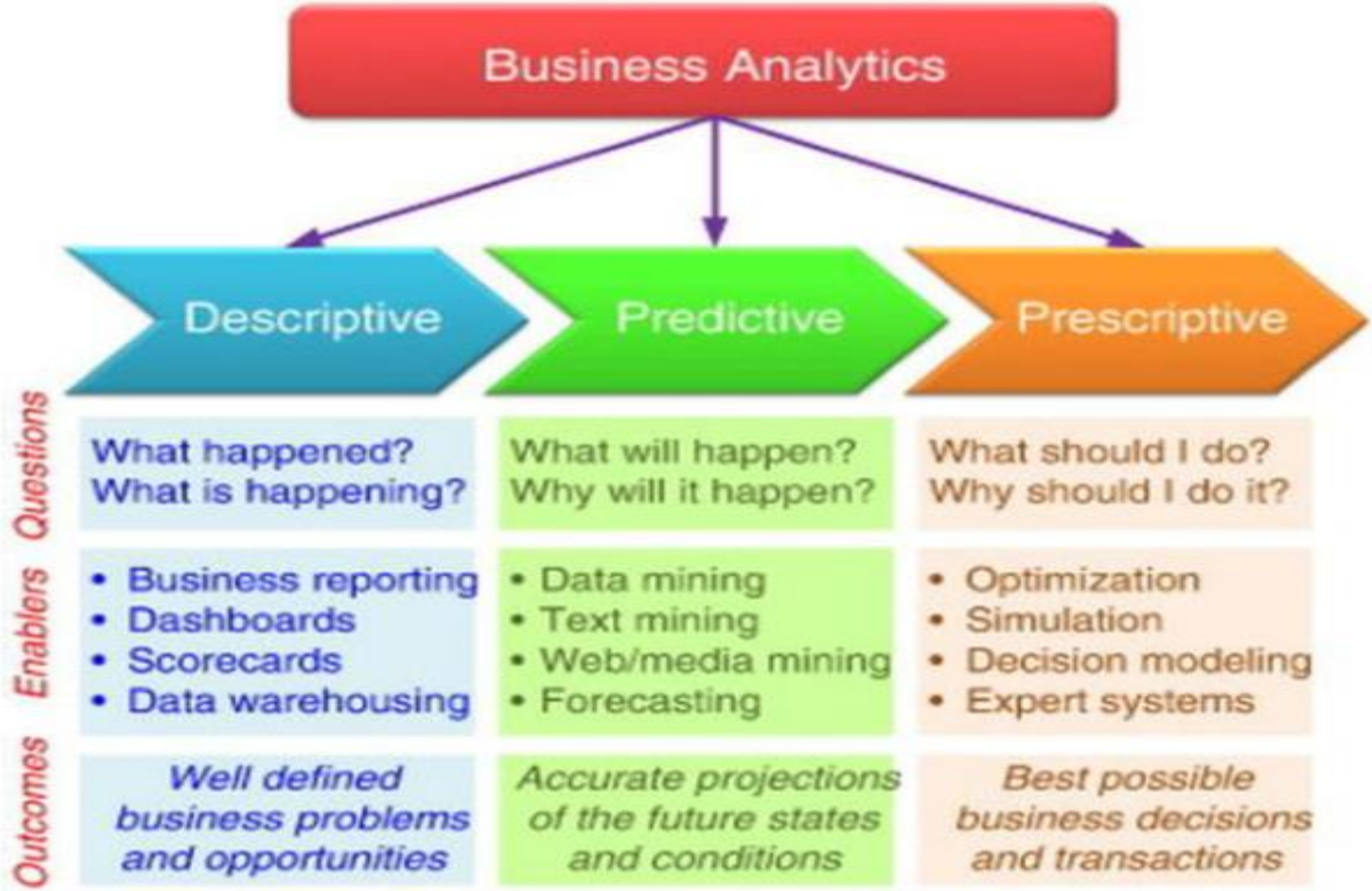
- Refers to the messiness or trustworthiness of the data.
- With many forms of big data, quality and accuracy are less controllable (just think of Twitter posts with hashtags, abbreviations, typos and colloquial speech as well as the reliability and accuracy of content) but big data and analytics technology now allows us to work with these type of data.
- The volumes often make up for the lack of quality or accuracy.

Value

Having access to big data is no good unless we can turn it into value.

Companies are starting to generate amazing value from their big data.

The Value of Data (alternative view)



The 7 Vs of Big Data

Validity

- The interpreted data having a sound basis in logic or fact – is a result of the logical inferences from matching data.

  Volume -Validity = Worthlessness

Visibility

- The state of being able to see or be seen – is implied

https://livingstoneadvisory.com/2013/06/vs-big-data/

The 10 Vs of Big Data

Variability

- Big data is also variable because of the multitude of data dimensions resulting from multiple disparate data types and sources.

Vulnerability

- Big data brings new security concerns. After all, a data breach with big data is a big breach

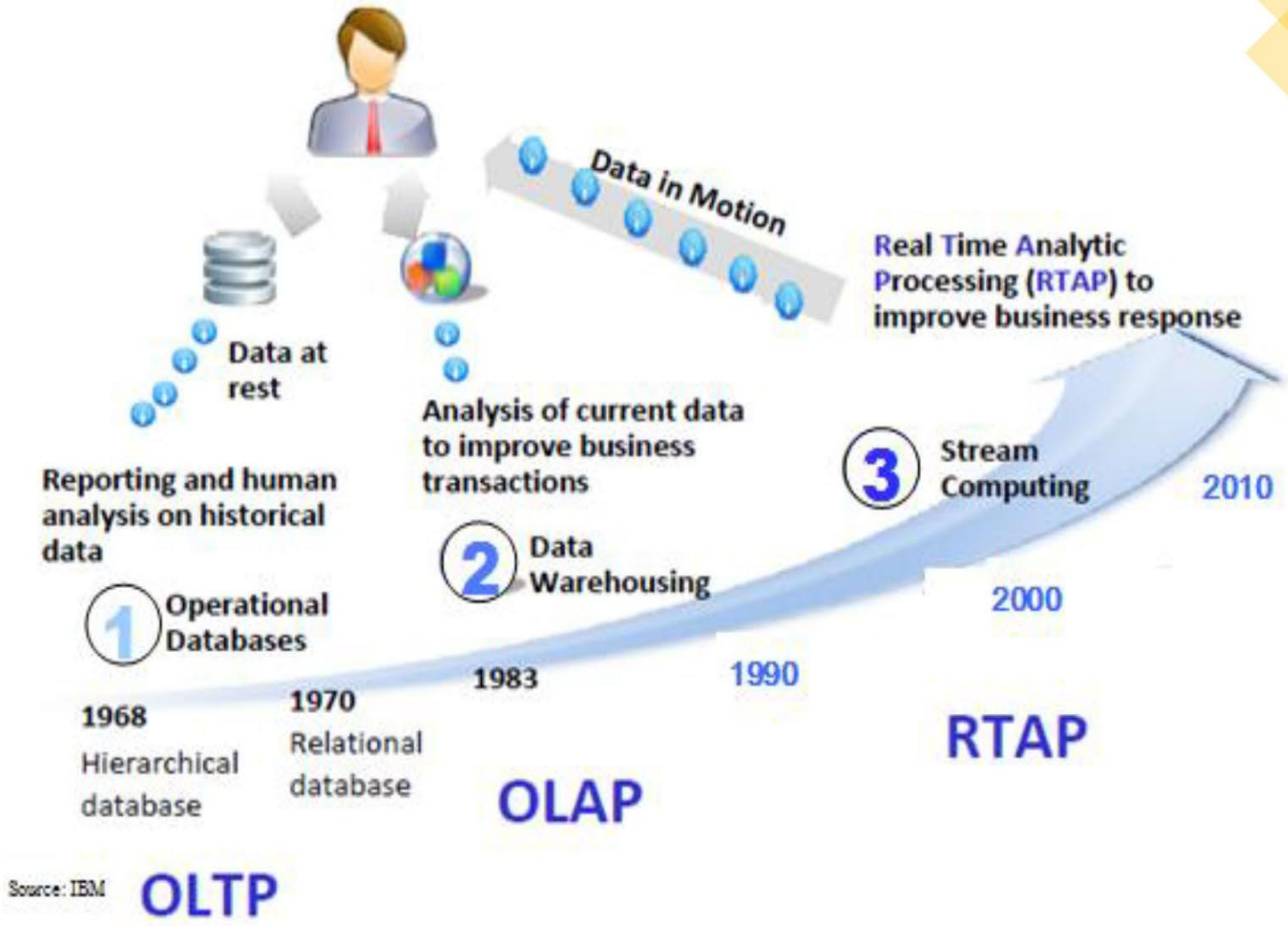
Volatility

- How old does your data need to be before it is considered irrelevant, historic, or not useful any longer? How long does data need to be kept for?

https://tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx

Harnessing Big Data

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



Challenges of Big Data

- ***Data Representation***
  - *Many datasets have certain* ***levels of heterogeneity in type****,* structure, semantics, organization, granularity, and accessibility.
  - Nevertheless, ***an improper data representation will reduce the value of the original data and may even obstruct effective data analysis.***
- ***Redundancy Reduction and Data Compression***
  - *Generally, there is a high level of redundancy in datasets.*
  - *Redundancy reduction and data compression is effective* to reduce the indirect cost of the entire system on the premise that the potential values of the data are not affected.
  - For example, ***most data generated by sensor networks are highly redundant, which may be filtered and compressed at orders of magnitude.***

Challenges of Big Data

- ***Data Life Cycle Management***
  - Generally speaking, ***values hidden in big data depend on data freshness.***
  - Therefore, an importance principle related to the analytical value should be developed ***to decide which data shall be stored and which data shall be discarded.***
- ***Analytical Mechanism***
  - The analytical system of big data shall process ***masses of heterogeneous data within a limited time.***
- ***Data Confidentiality***
  - Data contains details of ***the lowest granularity*** and ***some sensitive information such as credit card numbers***.
  - Therefore, analysis of big data ***may be delivered to a third party for processing only when proper preventive measures are taken*** to protect the sensitive data, to ensure its safety.

Challenges of Big Data

- ***Expendability and Scalability***
  - *The analytical system of* ***big data must support present and future datasets***.
  - The ***analytical algorithm*** must be able to ***process increasingly expanding and more complex datasets.***
- ***Cooperation***
  - *Analysis of big data is* ***an interdisciplinary research****, which requires* experts in different fields cooperate to harvest the potential of big data.
  - A comprehensive big data ***network architecture must be established to help scientists*** and ***engineers in various fields access different kinds of data*** and fully utilize their expertise, so as to cooperate to complete the analytical objectives.

Big Data Applications

- ***Application of Big Data in Enterprises:***
  - In particular, **on marketing**, with correlation analysis of big data, enterprises can more accurately predict the behavior of consumers and mine new business modes.
  - **On sales planning**, after comparison of massive data, ***enterprises can optimize their commodity prices.***
  - **On operation**, enterprises can improve their operation efficiency and operation satisfaction, optimize the input of labor force, accurately forecast personnel allocation requirements, ***avoid excess production capacity, and reduce labor cost.***
  - **On supply chain**, using big data, enterprises may ***conduct inventory optimization, logistic optimization, and supplier coordination***, etc., to mitigate the gap between supply and demand, control budgets, and improve services.

Big Data Applications

- ***Application of IoT based Big Data***
  - ***Trucks of UPS are installed with sensors, wireless adapters, and GPS, so the Headquarter can track truck positions and prevent engine failures.***
  - Meanwhile, this equipment also help UPS supervise and manage its employees, and optimize delivery routes.
  - The optimal delivery routes specified by UPS for trucks are derived from their past driving experience. ***In 2011, UPS drivers have driven for nearly 48.28 million km less***.
  - ***Smart city is a research area based on the application of Internet of Things data.***
  - For example, Department of Park Management of Dade County saved ***one million USD in water bills due to timely identifying and fixing water pipes that were running and leaking this year.***

Big Data Applications

- ***Application of Online Social Network-Oriented Big Data***
  - ***Content-Based Applications:*** *Language and text are two most important forms* of representation in SNS. Through the analysis of language and text, user preferences, emotions, interests, and demands, etc. may be revealed.
  - ***Structure-Based Applications:*** *On SNS with users as nodes, social relation*, interest, and hobbies, etc. aggregate relations among users into a clustered structure. Such structure with close relations among internal individuals but loose externally relations is also called a community.
  - The U.S. Santa Cruz Police Department experimented by applying data to conducting predictive analysis. *By analyzing SNS*, ***the police department can discover crime trends and crime modes, and even predict the crime rates in major regions.***

Big Data Applications

- ***Application of Online Social Network-Oriented Big Data***
  - ***In April 2013, Wolfram Alpha***, a computing and search engine of the U.S., ***studied the law of social behaviors of users*** by analyzing social data of more than one million American users of Facebook.
    - According to the analysis, it was found that **most users of Facebook fall in love in their early 20s**, get engaged when they are about 27 years old, get married when they are about 30 years old, and have slow changes in their marriage relationship between 30 and 60 years old. Such research results are highly consistent with the demographic census data of the U.S.
  - ***Global Pulse conducted a research*** that revealed some laws in social and economic activities using SNS data. This project utilized publicly available Twitter messages in English, Japanese, and Indonesian ***from July 2010 to October 2011***, ***to analyze topics related to food, fuel, housing, and loan***.

Big Data Applications

Applications of Healthcare and Medical Big Data

- Big data has unlimited potential for effectively storing, processing, querying, and analyzing medical data. The application of medical big data will profoundly influence the human health.
- ***For example, Aetna Life Insurance Company*** scanned 600,000 laboratory test results and 180,000 claims through a series of detection test results of metabolic syndrome of patients in three consecutive years. In addition, it summarized the final result into an extreme personalized treatment plan to assess the dangerous factors and main treatment plans of patients.
- The ***Mount Sinai Medical Center*** in the U.S. utilizes ***technologies of Ayasdi***, a big data company, to analyze all genetic sequences of Escherichia Coli, including over one million DNA variants, ***to know why bacterial strains resist antibiotics.***

Background of Data Analytics

- Big data analytics is the process of examining large amounts of data of a variety of types.
- The primary goal of big data analytics is to help companies make better business decisions.
- Analyze huge volumes of transaction data as well as other data sources that may be left untapped by conventional business intelligence (BI) programs.

Data Analytics

Big data Consist of

- Uncovered hidden patterns.
- Unknown correlations and other useful information.
- Such information can provide business benefits.
- More effective marketing and increased revenue.

Data Analytics

- Big data analytics can be done with the software tools commonly used as part of advanced analytics disciplines such as **predictive analysis** and **data mining.**

- But the unstructured data sources used for big data analytics may not fit in traditional data warehouses.

- Traditional data warehouses may not be able to handle the processing demands posed by big data.

Data Analytics

- The technologies associated with big data analytics include NoSQL technologies associated with big data analytics include NoSQL databases, Hadoop and MapReduce.
- Knowledge about these technologies form the core of an open-source software framework that supports the processing of large data sets across clustered systems.
- Big Data analytics initiatives include
  - Internal data analytics skills
  - High cost of hiring experienced analytics professionals,
  - Challenges in integrating Hadoop systems and data warehouses

Data Analytics

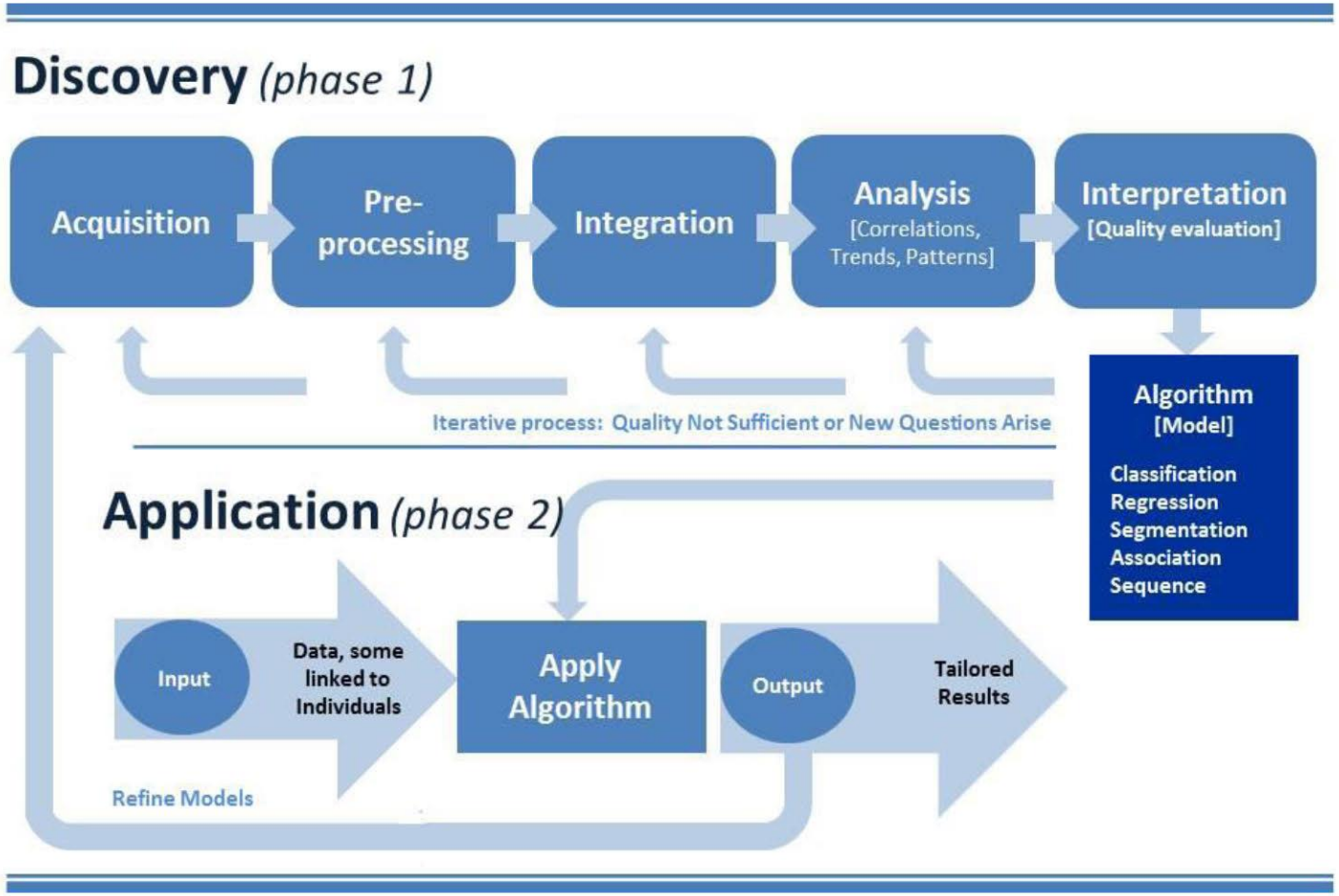
- Big Analytics delivers competitive advantage compared to the traditional analytical model.
- Big Analytics describes the efficient use of a simple model applied to volumes of data that would be too large for the traditional analytical environment.
- Research suggests that a simple algorithm with a large volume of data is more accurate than a sophisticated algorithm with little data.

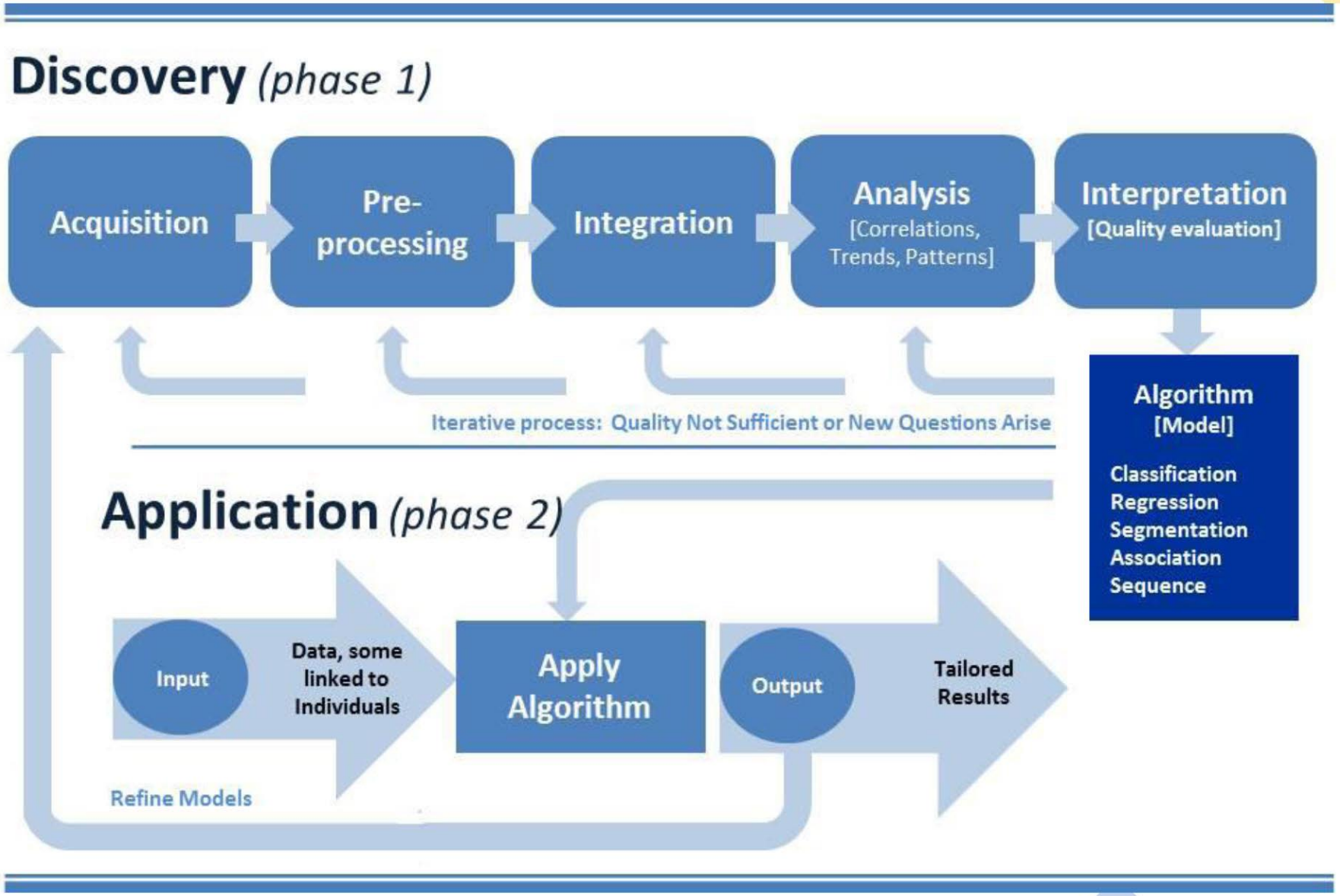
Data Analytics

- ***Data analytics is broken down into four basic types.***

1. ***Descriptive analytics:*** This describes what has happened over a given period of time. Have the number of views gone up? Are sales stronger this month than last?
2. ***Diagnostic analytics:*** This focuses more on why something happened. This involves more diverse data inputs and a bit of hypothesizing. Did the weather affect beer sales? Did that latest marketing campaign impact sales?
3. ***Predictive analytics:*** This moves to what is likely going to happen in the near term. What happened to sales the last time we had a hot summer? How many weather models predict a hot summer this year?
4. ***Prescriptive analytics:*** This suggests a course of action. If the likelihood of a hot summer is measured as an average of these five weather models is above 58%, we should add an evening shift to the brewery and rent an additional tank to increase output.

The Process of Data Analytics




Discovery (phase 1)
Acquisition
Pre-processing
Integration
Analysis
[Correlations, Trends, Patterns]
Interpretation
[Quality evaluation]
Algorithm
[Model]
Classification
Regression
Segmentation
Association
Sequence
Iterative process: Quality Not Sufficient or New Questions Arise
Application (phase 2)
Input
Data, some linked to Individuals
Apply Algorithm
Output
Tailored Results
Refine Models

Data Analytics Process: Discovery

The knowledge discovery phase involves

- gathering data to be analyzed.
- pre-processing it into a format that can be used.
- consolidating it for analysis,
- analyzing it to discover what it may reveal.
- and interpreting it to understand the processes by which the data was analyzed and how conclusions were reached.

Data Analytics Process: Discovery

Acquisition

- Data acquisition involves collecting or acquiring data for analysis.
- Acquisition requires access to information and a mechanism for gathering it.

Pre-processing

- Pre-processing is necessary if analytics is to yield trustworthy , useful results.
- places it in a standard format for analysis.

Data Analytics Process: Discovery

Integration

- Integration involves consolidating data for analysis.
- Retrieving relevant data from various sources for analysis
- Eliminating redundant data or clustering data to obtain a smaller representative sample.

Analysis

- Searching for relationships between data items in a database or exploring data in search of classifications or associations.
- Analysis can yield descriptions or predictions.
- Analysis based on interpretation, organizations can determine whether and how to act on them.

Data Analytics Process: Discovery

Interpretation

- Analytic processes are reviewed by data scientists to understand results and how they were determined.
- Interpretation involves retracing methods, understanding choices made throughout the process and critically examining the quality of the analysis.
- It provides the foundation for decisions about whether analytic outcomes are trustworthy.

Data Analytics Process: Application

Application

- Associations discovered amongst data in the knowledge phase of the analytic process are incorporated into an algorithm and applied.
- In the application phase organizations gather the benefits of knowledge discovery.
- Through application of derived algorithms, organizations make determinations upon which they can act.

Role of Distributed System in Big Data

What is a Distributed System?

- Consists of a collection of autonomous computers, connected through a network and distribution middleware
- Enables computers to coordinate their activities and to share the resources of the system
- Users perceive the system as a single, integrated computing facility.

Big data is distributed data

BIG DATA IS **DISTRIBUTED DATA** : DATA IS SO MASSIVE IT CANNOT BE STORED OR PROCESSED BY A SINGLE NODE.

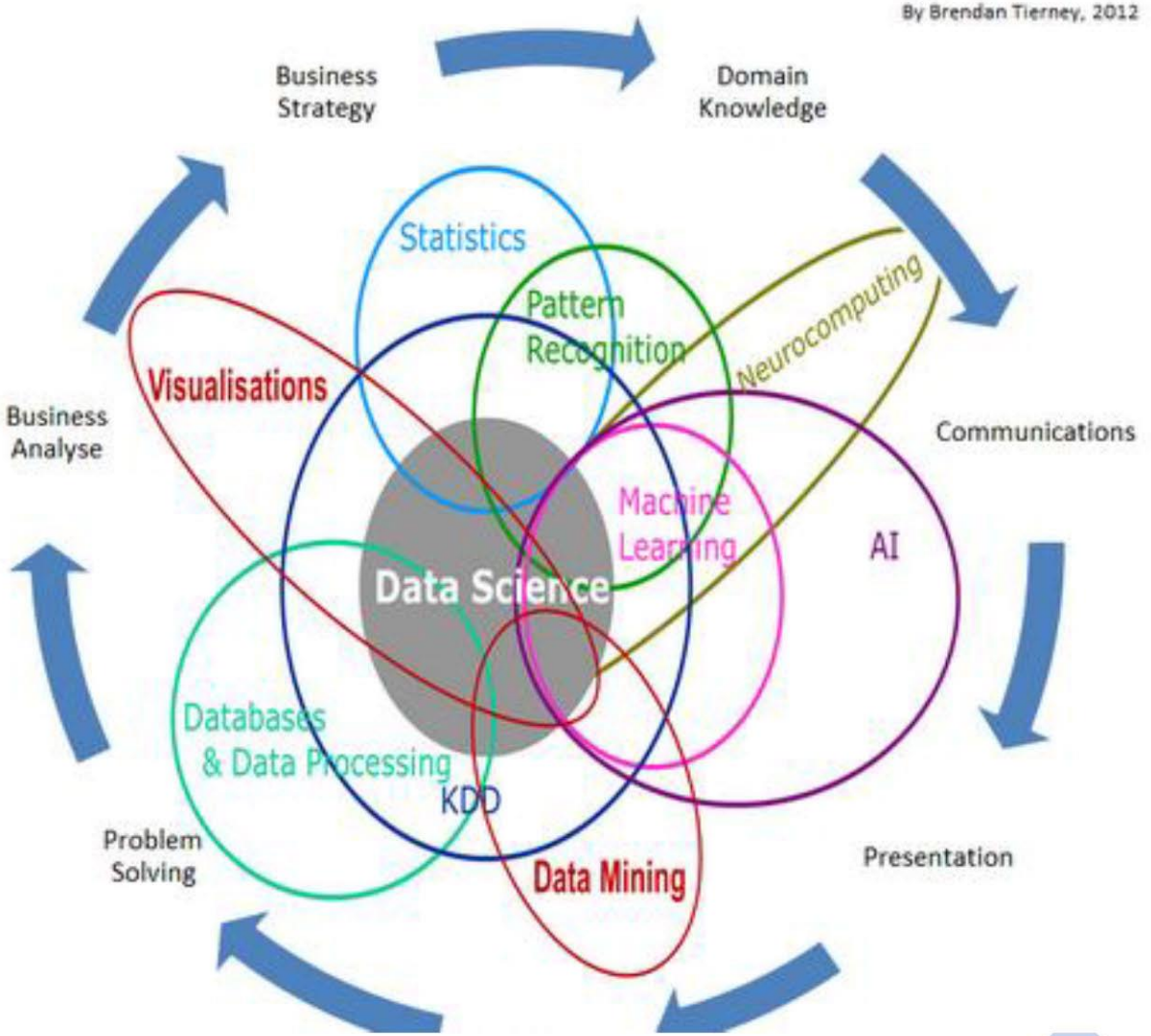
THE WAY TO SCALE FAST AND AFFORDABLY IS TO USE COMMODITY HARDWARE TO DISTRIBUTE THE STORAGE AND PROCESSING OF OUR MASSIVE DATA STREAMS ACROSS SEVERAL NODES, ADDING AND REMOVING NODES AS NEEDED.

Distributed data generation is fueling big data growth

- The reason we have data problems so big that we need large-scale distributed computing architecture to solve is that the creation of the data is also large-scale and distributed.

- Most of us walk around carrying devices that are constantly pulsing all sorts of data into the cloud and beyond – our locations, our photos, our tweets, our status updates, our connections, even our *heartbeats*.

Data Science Is Multidisciplinary



Who are Data Scientist?

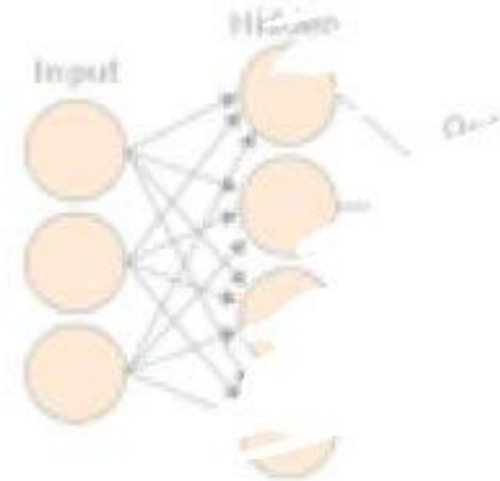
- High ranking professional with training and curiosities to make discovery in the world of big data.
- The people who understand how to fish out answers to important business questions from today's tsunami of unstructured information.
- Newly coined term , in 2008 by D.J Patil and Jeff Hammerbacher.
- A hybrid of data hacker, analyst, communicator, and trusted adviser. The combination is extremely powerful and rare.

Data Scientist

- Sudden appearance of Data Scientist on the business scene reflects the fact that companies are now wrestling with information that comes in varieties and volumes never encountered before.

- If the organization stores multiple petabytes of data, if the information most critical to the business resides in forms other than rows and columns of numbers, or if answering the biggest question would involve a "mashup" of several analytical efforts, it has got a big data opportunity.

The Data Scientist



- A New Role Exists – the **Data Scientist**
  - One Part Scientist/Statistician
  - Two Parts Sleuth/Artist
  - One Part Programmer
  - Focused on *data* not models
- Working with **analysts** to create business value

Data scientist: a brand-new profession

- Data Scientist: The Sexiest Job of the 21st Century [Harward Business Review 2013]
- Data scientist? A guide to 2015's hottest profession [Mashable 2015]
- “It’s official – data scientist is the best job in America” [Forbes, 2016]
- "This hot new field promises to revolutionize industries from business to government, health care to academia."
  - — *The New York Times*

Successful Data Scientist Characteristics

Intellectual curiosity, Intuition

- Find needle in a haystack(something that is difficult to locate in a much larger space)
- Ask the right questions – value to the business

Communication and engagements

Presentation skills

- Let the data speak but tell a story
- Storyteller – drive business value not just data insights

Creativity

- Guide further investigation

Business Savvy

- Discovering patterns that identify risks and opportunities
- Measure

Role/Skill of Data Scientist

Data Scientist should have skill set to

- use technologies that make taming big data possible, including Hadoop (the most widely used framework for distributed file system processing) and related open-source tools, cloud computing, and data visualization.
- make discoveries while swimming in pool of data
- bring structure to large quantities of formless data and make analysis possible
- identify rich data sources, join them with other, potentially incomplete data sources, and clean the resulting set

Role/Skill of Data Scientist

Data Scientist should have skill set to…(Contd)

- communicate what they've learned and suggest its implications for new business directions.
- be creative in displaying information visually and making the patterns they find clear and compelling.
- fashion their own tools and even conduct academic-style research.
- write code.
- desire to go beneath the surface of a problem, find the questions at its heart, and distill them into a very clear set of hypotheses that can be tested.

Data Scientist Job Description

- Amazon's Shopper Marketing & Insights team focuses on serving the advertisers and our overall ad business to provide strategic media planning, customer insights, targeting recommendations, and measurement and optimization of advertising.

- We are hiring outstanding Data Scientists who will use innovative statistical and machine learning approaches to drive advertising optimization and contribute to the creation of scalable insights. The ideal candidate should have one hand on the white-board writing equations and one hand on the keyboard writing code.

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21th century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning: decision trees, random forests, logistic regression
- Unsupervised learning: clustering, dimensionality reduction
- Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing package e.g. R
- Databases SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Pig
- Custom reducers
- Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

THE STATE OF THE DATA SCIENTIST

HIGHEST EDUCATION LEVEL

TOP PRIMARY SKILLS

1. Data Analysis
2. R
3. Python
4. Data Mining
5. Machine Learning

INDUSTRY GROWTH

TOP EDUCATIONAL BACKGROUNDS

1. Computer Science
2. Business Admin
3. Statistics
4. Mathematics
5. Physics

TOP INDUSTRIES EMPLOYING DATA SCIENTISTS

1. Information Technology & Services
2. Internet
3. Computer Software
4. Education
5. Banking & Financial Services

HealthCatalyst
ignite outcomes improvement

Data analyst vs. data scientist

Data analyst

EDUCATION

- Four-year degree in mathematics, statistics or business with a focus on analytics.

EXPERIENCE AND SKILLS

- Background in mathematics and statistics with a solid understanding of data mining techniques.
- Companies seek candidates with strong written and verbal communication skills, as well as familiarity with data mining, data modeling, R, SQL, statistical analysis, database management and data analysis.

EXPECTATIONS

- Designing and maintaining data systems and databases using statistical tools to interpret data, prepare reports that can communicate trends, patterns and predictions based on data, as well as conduct consumer data research and analytics.
- Work with customer-centric algorithm models and be able to shape them to each customer as required. Extract actionable insights from large databases, help translate data into visualizations, metrics and goals, and have a general proactive approach.

Data scientist

EDUCATION

- Masters or doctorates in mathematics, statistics or computer science.

EXPERIENCE AND SKILLS

- Combination of mathematical and statistical knowledge, programming expertise, as well as analytical skills including familiarity with machine learning, software development, Hadoop, Java, data mining/data warehouse, data analysis and Python.
- Experience working and creating data architectures; familiarity with advanced statistical techniques/concepts.

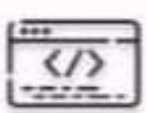
EXPECTATIONS

- Data scientist roles require the creation of algorithms and predictive models that extract the needed information to solve complex problems. Focus on designing and constructing new processes for data modeling and production.
- Able to use predictive modeling to increase and optimize customer experiences, revenue generation and ad targeting. Develop custom data models and algorithms; develop processes and tools to monitor and analyze model performance.
- Ask unique questions and predict future trends.

Data Scientist

uses statistics and machine learning to make predictions and answer key business questions

Skills - Math, Programming, Statistics

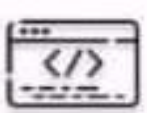


Tech - SQL, Python, R, Cloud

Data Engineer

build and optimize the systems that allow data scientists and analysts to perform their work

Skills - Programming, BigData & Cloud



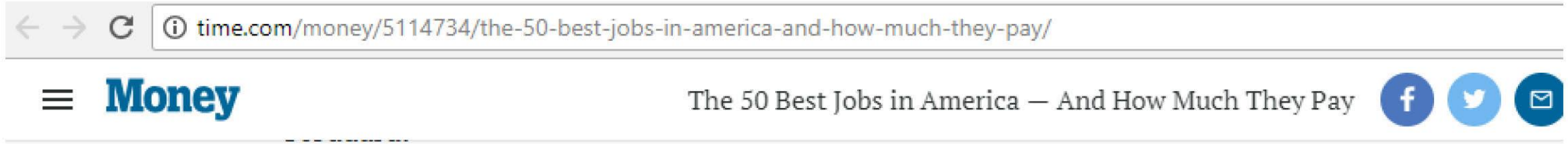
Tech - SQL, Python, Cloud, Distributed Computing

Data Analyst

deliver value by taking data, communicating the results to help make business decisions

Skills - Communication, Business Knowledge

Tech - SQL, Excel, Tableau



Looking for a new, more promising gig? Here's Glassdoor's full list of the 50 best jobs in the U.S. for 2018, including links to open positions.

1. Data Scientist

- Job Score: 4.8
- Job Satisfaction Rating: 4.2
- Number of Job Openings: 4,524
- Median Base Salary: $110,000

The 50 best jobs in America

Thank You

???