

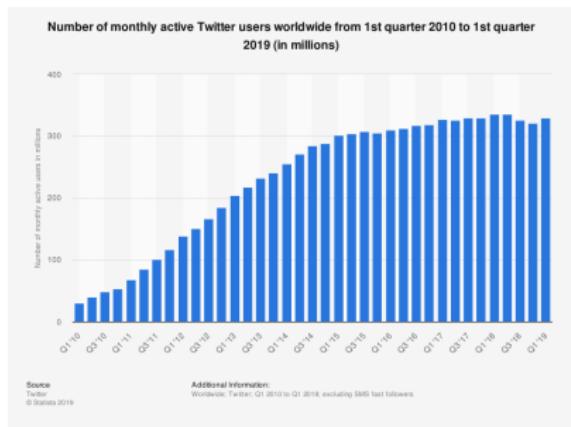
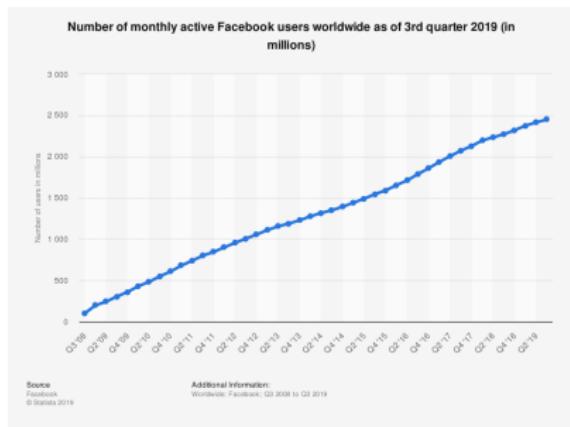
INTRODUCTION TO BIG DATA ANALYTICS

■ Imdadullah Khan

Big Data Generation and Growth

- Data has been generated at an exploding rate in recent years
- Organizations collect trillions of bytes of information about their customers, suppliers, and operations every day
- Large pools of data is being captured, communicated, aggregated, stored, and analyzed by all kinds of organizations, businesses, academia, governments
- Multimedia and individuals with smartphones on social network sites are continuously fueling the exponential growth of data

Big Data Generation and Growth



Big Data Generation and Growth

Where data comes from?

- Internet users generate about 2.5 quintillion bytes of data each day¹
- In 2018, internet users spent 2.8 million years online²
- Social media accounts for 33% of the total time spent online ²
- In 2019, there are 2.3 billion active Facebook users
- Twitter users send nearly half a million tweets every minute¹
- By 2020, every person will generate 1.7 megabytes in just a second¹
- By 2020, there will be 40 trillion gigabytes of data (40 zettabytes)³
- 90% of all data has been created in the last two years ⁴

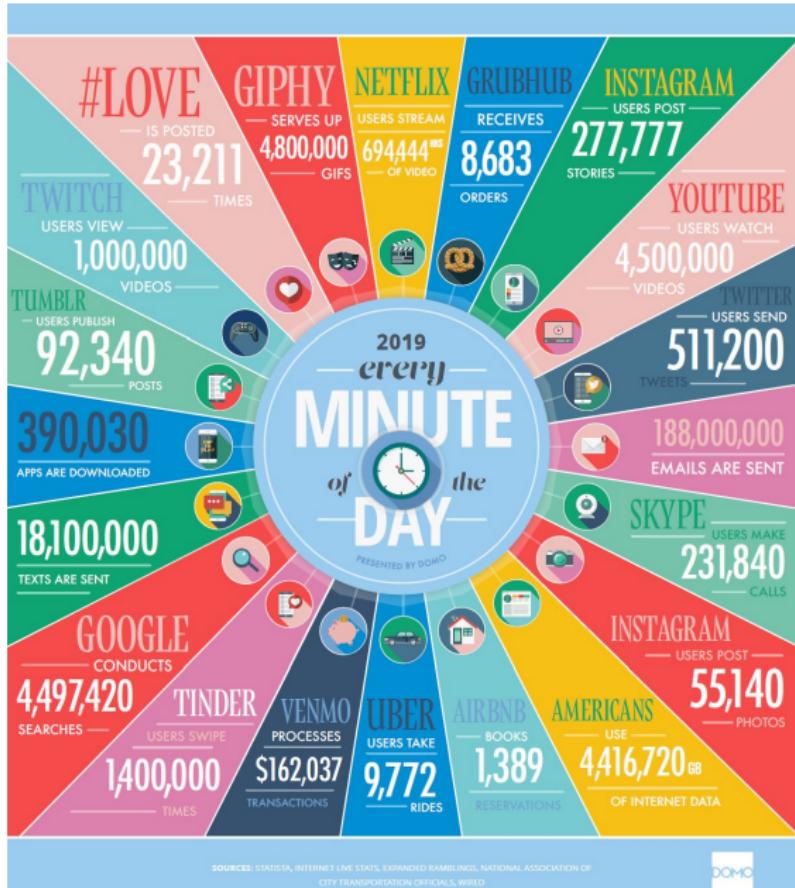
¹ Domo report (a company with data analytic platform for businesses)

² Global Web Index report (a company with big data analytic platform)

³ EMC (Dell EMC provides big data solutions)

⁴ IBM

Big Data Generation and Growth



Big Data Generation and Growth



If the Digital Universe were represented by the memory in a stack of tablets, in 2013 it would have stretched two-thirds the way to the Moon*

By 2020, there would be 6.6 stacks from the Earth to the Moon*

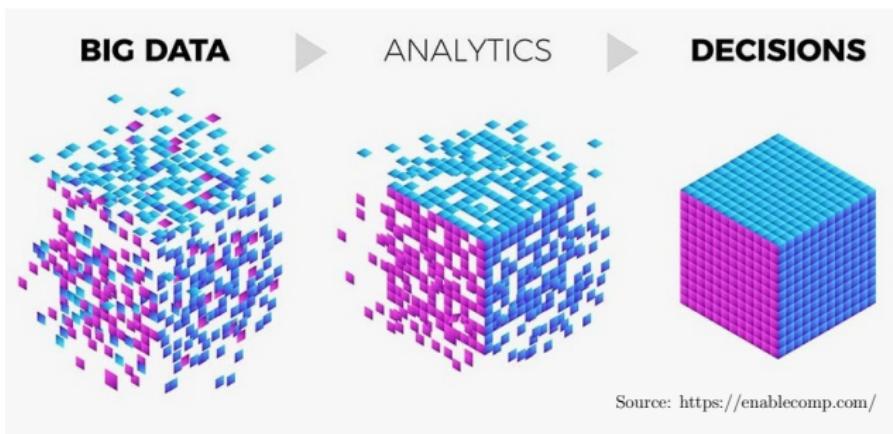
- 90% of all data has been created in the last two years⁵

What is Big Data

- “**Big data**” refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze
- As technology advances over time, the size of datasets that qualify as big data will also increase
- The definition can vary by sector, depending on what kinds of software tools are commonly available and what sizes of datasets are common in a particular industry
- With those caveats, big data in many sectors today will range from a few dozen terabytes to multiple petabytes (thousands of terabytes)

Big Data Analytics

- **Data:** Set of values of qualitative or quantitative variables
- **Information:** Meaningful or organized data
- **Data Analytics:** The process of examining data in order to draw and communicate useful conclusions about the information it contains



Then and Now

- Data Analytics has been around for years
- Even in 1950's, businesses were using basic analytics (manual examination) on data (essentially numbers in a spreadsheet) to uncover insights and trends
- New tools and technologies bring speed and efficiency in techniques
- Today, businesses analyze data and can identify insights for immediate decisions
- The ability to work faster and stay agile gives organizations a competitive edge they did not have before

Why is Big Data Analytics Important

Organizations analyze data

- to identify new opportunities
- to gain insights that lead to smarter business decisions
- to identify methods for more efficient operations
- to maximize larger revenues and higher profits
- to keep customers satisfied

Top three factors businesses got
the most value in

- Cost reduction
- Faster, better decision making
- New products and services



Why enterprises use Big Data Analytics

Companies are using big data analytics for all types of decisions

The Evolution of Decision Making: How Leading Organizations Are Adopting a Data-Driven Culture

A REPORT BY HARVARD BUSINESS REVIEW ANALYTIC SERVICES

Sponsored by



What enterprises use Big Data Analytics for

- Competitor Analysis
 - Online traffic to websites and related social media
- Market Analysis
 - Trends and market segment analysis
- Productivity Enhancement
 - Analyze employees tracking data
- Cost Cutting
 - Reduce energy bills, optimize routes, predict demands, process efficiency and automation⁶
- Targeted Marketing
 - Analyze purchasing history and target the right people for a product
- Improved Customer Relations
 - Analyze customer feedback and make adjustments

⁶ Forbes (01/08/2016) Big Data Analytics' Potential to Revolutionize Manufacturing Is Within Reach

Industries Benefiting from Big Data Analytics

- **Retail:** Advertising, Targeted marketing, recommendation system, customer loyalty, inventory management, demand prediction, real-time pricing
- **Banking and Financial:** Customer loyalty and churn, fraud detection, risk assessment
- **Brands:** 66% brands use data analytics for product and service launch, appropriate timings
- **Logistics and Transportation:** Fleet management, maintenance needs, drivers risk assessment, real time tracking
- **Health Care:** Efficiency in healthcare operations, predictive analytics, outbreak prediction, immunization strategy

Google's AI system can beat doctors at detecting breast cancer

By Hanna Zlady, **CNN BUSINESS** January 2, 2020

- **Government and Large Utility Companies:** Surveys and census, planning and development, public health, education, energy supply & demand management

36,624 views | Jan 13, 2016, 02:24am

Big Data Facts: How Many Companies Are Really Making Money From Their Data?



Bernard Marr Contributor (0)
Enterprise Tech

Forbes

FORTUNE

For the airline industry, big data is cleared for take-off

BY KATHERINE NOYES

June 19, 2014 8:10 PM EST

FORTUNE

How commercial insurer FM Global uses data science to reduce client risk

BY HEATHER CLANCY
December 10, 2014 2:00 AM EST

How Big Data is reducing costs and improving performance in the upstream industry



By BINU MATHEW, GLOBAL HEAD OF DEVELOPMENT & PRODUCT MANAGEMENT, GE OIL & GAS DIGITAL on 12/13/2016

FORTUNE

Cropping up on every farm: Big data technology

BY KATHERINE NOYES

May 30, 2014 11:00 PM EST

Industries Benefiting from Big Data Analytics

FORTUNE Bright lights, big cities, bigger data

BY SHALENE GUPTA

October 31, 2014 3:42 AM EST

Industries Benefiting from Big Data Analytics

FORTUNE

Can Big Data cure cancer?

BY MIGUEL HELFT

July 24, 2014 4:31 PM EST

Can smart sensor systems anticipate and avoid danger?

Kate Pisa, CNN

⌚ Updated 1508 GMT (2308 HKT) January 21, 2020



FORTUNE At Coca-Cola Bottling, flash memory energizes big data efforts

BY KATHERINE NOYES
June 28, 2014 12:25 AM EST

Will big data help end **FORTUNE** discrimination—or make it worse?

BY KATHERINE NOYES
January 16, 2015 1:16 AM EST

Industries Benefiting from Big Data Analytics

Fitness app that revealed military bases highlights bigger privacy issues



by Selena Larson [@selenalarson](#)

January 29, 2018: 5:23 PM ET

FORTUNE

What's on trend this season for the fashion industry? Big data

BY KATHERINE NOYES

September 22, 2014 5:26 PM EST

FORTUNE How GE generates \$1 billion from data

BY HEATHER CLANCY
October 11, 2014 1:16 AM EST

FORTUNE Police are crunching data to stop murders before they happen

BY SHALENE GUPTA

February 9, 2015 7:00 PM EST

FORTUNE

Predictive analytics, a potent prescription for health care

BY HEATHER CLANCY

January 6, 2015 12:03 AM EST

Big Data Analytics - Market

- 12% - the rate of increase for big data and business analytics use from 2018 to 2019 ⁷
- \$189.1 billion – projected worldwide revenues for big data and business analytics solutions for 2019 ⁷
- \$274.3 billion – projected worldwide revenues for big data and business analytics solutions by 2022 ⁷
- 13.2% - projected compound annual growth rate (CAGR) of big data and business analytics within the five-year period, 2018-2022 ⁷

⁷ International Data Corporation (IDC) - Big data analytics company

Big Data Analytics - Market

Big Data & Business Analytics Solutions Worldwide Revenues

(Projected in US\$ B, 2019-2022)



Source: IDC



FinancesOnline
REVIEWS FOR BUSINESS

Sources of Big Data



Information Processing & Management

Volume 54, Issue 5, September 2018, Pages 758-790



A survey towards an integration of big data analytics to big insights for value-creation

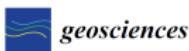
Mandeep Kaur Saggi , Sushma Jain

Sources: Machine Generated Data

- Biggest source of big data
- Temperature sensors, GPS navigator, Satellite imagery, Apps,
- Increasing number of smart devices, IoT
- A 12 hours flight produces 84TB of data, sensors, temperature, pressure, accelerometer, turbulence
- Smart City, Smart Transportation
- Think about the volume of video data collected at Lahore Safe City Authority Control Room
- Generally, such data is unstructured

Sources: People Generated Data

- Blogs, social network posts, keywords search, photo sharing, pictures, emails, ratings and reviews
- Daily facebook data 30+ PB > All US Academic libraries (2 PB)
- Companies use 12PB/day Twitter data for sentiment analyses around their products
- Could be used for disaster management, e.g. to identify and measure affected areas and channel resources



Review

Big Data in Natural Disaster Management: A Review

Manzhu Yu ^{*}, Chaowei Yang and Yun Li

NSF Spatiotemporal Innovation Center, George Mason University, 4400 University Drive, Fairfax, VA 22030, USA; cyang3@gmu.edu (C.Y.); yli380@gmu.edu (Y.L.)

* Correspondence: myu7@gmu.edu

Received: 12 March 2018; Accepted: 3 May 2018; Published: 5 May 2018

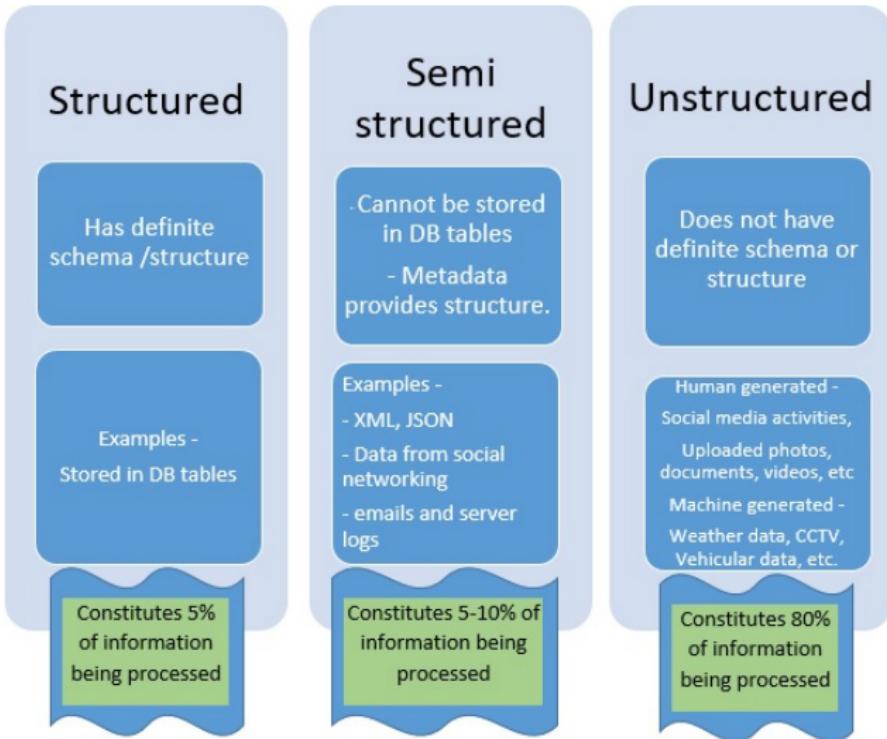


- Typically unstructured, or at best semi-structured such as emails, where the header has somewhat of a structure, except in few cases such as filling up a survey form
- Generally more text: 500 million tweets per day

Sources: Organization Generated Data

- LUMS Students Data, ESPN Cricinfo, TCS tracking and shipment data
- Governments open data, Stock Records, Banks, e-Commerce
- Medical Records
- Optimize routes and optimal scheduling can save 50m by reducing each drivers route by one mile
- Combine Walmart sales data with Twitter sentiment analyses or events to launch a new product
- Estimate demands
- Fraud Detection
- Highly Structured Data

Categories of Data



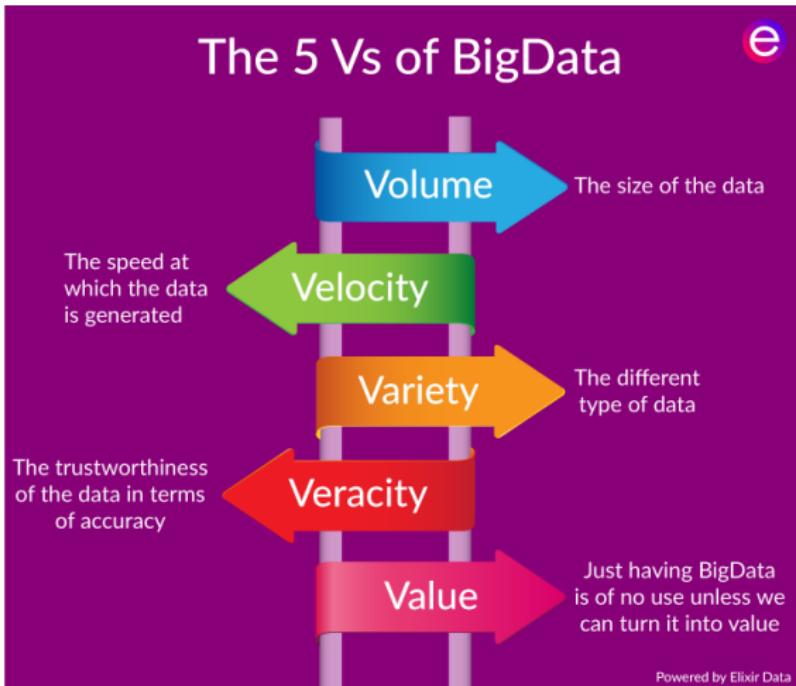
Data Definition Framework

Data Format	
Data Source	Structured
	Unstructured
Internal	<p>Structured</p>  <p>Human-Generated</p> <ul style="list-style-type: none">• Survey ratings• Aptitude testing <p>Machine-Generated</p> <ul style="list-style-type: none">• Web metrics from Web logs• Product purchase from sales Records• Process control measures <p>Unstructured</p>  <p>Human-Generated</p> <ul style="list-style-type: none">• Emails, letters, text messages• Audio transcripts• Customer comments• Voicemails• Corporate video/communications• Pictures, illustrations• Employee reviews
External	 <p>Structured</p>  <p>Human-Generated</p> <ul style="list-style-type: none">• Number of Retweets, Facebook likes, Google Plus +1s• Ratings on Yelp• Patient ratings <p>Machine-Generated</p> <ul style="list-style-type: none">• GPS for tweets• Time of tweet/updates/postings <p>Unstructured</p>  <p>Human-Generated</p> <ul style="list-style-type: none">• Content of social media updates• Comments in online forums• Comments on Yelp• Video reviews• Pinterest images• Surveillance video



Aspects of Big: The 5 V's

- 1 Volume
- 2 Velocity
- 3 Variety
- 4 Veracity
- 5 Value



Aspects of Big: The 5 V's – Volume

Volume: size, scale, dimensionality,

- 204m emails/minute, if an email is 100KB, and see the volume

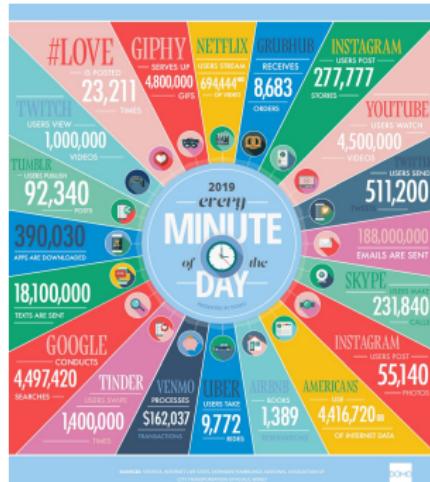


- Challenges: Acquisition, Storage, Retrieval, Processing Time
- Large dimensional data has more information, it is a blessing
- It is also a big curse, dealing with large dimensions is a core topic in this course

Aspects of Big: The 5 V's – Velocity

Velocity: Speed of data is very high

- Number of emails, twitter messages, photos, videos etc. per second



- Late decisions implies missed opportunities
- Real time processing vs Batch Processing (end of the day)

Aspects of Big: The 5 V's – Variety

Variety: Structural variety, different formats, models



- Medium variety, audio, text, video,
- DBMS, files, traffic logs, XML, code
- Online vs Offline,
- Real time vs Intermittent data (another way data varies)
- Challenges, requirement of specific analytics, Semantic, how to interpret

Aspects of Big: The 5 V's – Veracity

Veracity: quality of data

- Quality of Data itself has many issues (biases, anomalies, inconsistent measurements and units, incomplete and duplicate records)
- Volatility in data, updated/outdated, changing trends and sentiments
- Trustworthiness and reliability of sources and generation/processing
- Fake news, rumours, fake likes, fake followers

sci²force

Source: <https://datafloq.com/>

Sources of Data Veracity



Statistical biases



Lack of data lineage



Software bugs



Noise



Abnormalities



Information Security



Untrustworthy data sources



Falsification



Uncertainty and ambiguity of data



Duplication of data



Out of date and obsolete data



Human error

Aspects of Big: The 5 V's – Value

Value: Data can be turned into big value

- Data having no value is of no good to the company
- Should be able to meet strategic objectives
- Should amplify other technology innovations

5 Vs of Big Data: Value

The Economist Intelligence Unit report on surveying 476 executives

- 60% feel that data is generating revenue within their organizations
- 83% say it is making existing services and products more profitable
- 63% executives based in Asia said they are routinely generating value from data
- In the US, the figure was 58% and in Europe, 56%

26,624 views | Jan 13, 2016, 02:24am

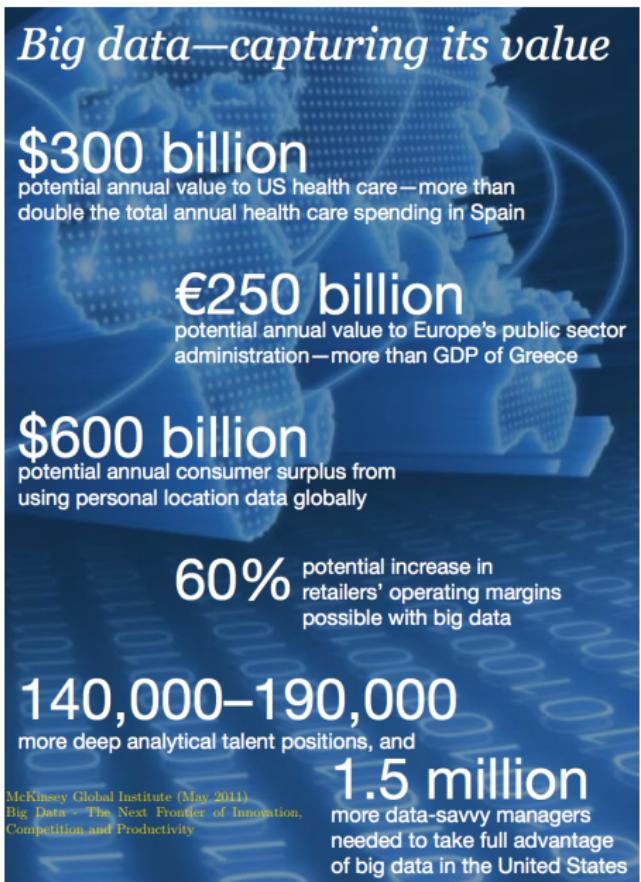
Big Data Facts: How Many Companies Are Really Making Money From Their Data?



Bernard Marr Contributor @
Enterprise Tech



5 Vs of Big Data: Value



5 Vs of Big Data: Value

European Union
public sector



United States
health care



Manufacturing



United States
retail



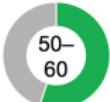
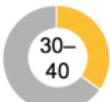
Location-based data



5 Vs of Big Data: Value

Exhibit E1

There has been uneven progress in capturing value from data and analytics

	Potential impact: 2011 research	Value captured %	Major barriers
Location-based data	<ul style="list-style-type: none">▪ \$100 billion+ revenues for service providers▪ Up to \$700 billion value to end users	 50–60	<ul style="list-style-type: none">▪ Penetration of GPS-enabled smartphones globally
US retail¹	<ul style="list-style-type: none">▪ 60%+ increase in net margin▪ 0.5–1.0% annual productivity growth	 30–40	<ul style="list-style-type: none">▪ Lack of analytical talent▪ Siloed data within companies
Manufacturing²	<ul style="list-style-type: none">▪ Up to 50% lower product development cost▪ Up to 25% lower operating cost▪ Up to 30% gross margin increase	 20–30	<ul style="list-style-type: none">▪ Siloed data in legacy IT systems▪ Leadership skeptical of impact
EU public sector³	<ul style="list-style-type: none">▪ ~€250 billion value per year▪ ~0.5% annual productivity growth	 10–20	<ul style="list-style-type: none">▪ Lack of analytical talent▪ Siloed data within different agencies
US health care	<ul style="list-style-type: none">▪ \$300 billion value per year▪ ~0.7% annual productivity growth	 10–20	<ul style="list-style-type: none">▪ Need to demonstrate clinical utility to gain acceptance▪ Interoperability and data sharing

1 Similar observations hold true for the EU retail sector.

2 Manufacturing levers divided by functional application.

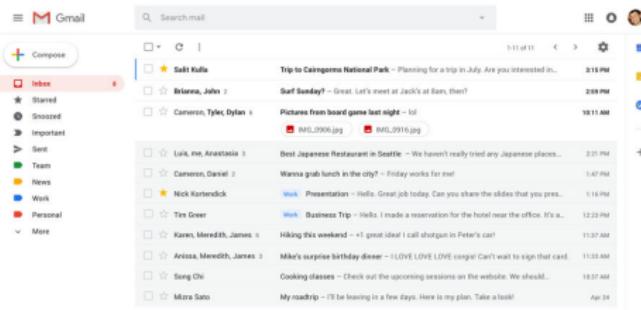
3 Similar observations hold true for other high-income country governments.

Types of Data

- Relational Data
- Text Data
- Multimedia Data
- Time Series Data
- Sequential Data
- Streams
- Graphs and Homogeneous Networks
- Graphs and Heterogeneous Networks

Types of Data: Text

- blogs, webpages, tweets, documents, emails
- High dimensionality, vocabulary, information retrieval, natural language processing
- Latest search engine for Walmart.com uses text analysis, machine learning and even synonym mining to produce relevant search results. Wal-Mart says adding semantic search has improved online shoppers completing a purchase by 10% to 15%. "In Wal-Mart terms, that is billions of dollars,"



Types of Data: Multimedia

- image, audio, video
- 'Fast food and video' company is training cameras on drive-through lanes to determine what to display on its digital menu board. When the lines are longer, the menu features products that can be served up quickly; when the lines are shorter, the menu features higher-margin items that take longer to prepare



Here's why some McDonald's restaurants are putting cameras in their dumpsters

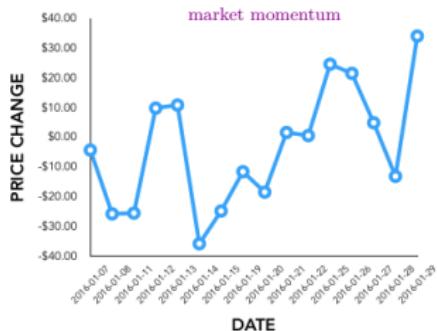


By Rachel Metz, CNN Business
Updated 1736 GMT (0136 HKT) December 18, 2020



Types of Data: Time Series

- Sequence of data points at equally spaced time intervals
- Sensor data, Stock market data, Forex rates, Temporal tracking (GPS), Smart Meters Data (AMI)
- Understanding the underlying forces and structure of observed data and fit a model to forecast, monitor or control
- Economic Forecasting, Sales Forecasting, Stock Market Analysis, Yield Projections, Process and Quality Control, Inventory Studies, Workload Projections, Census Analysis



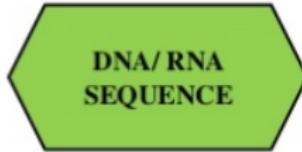
Types of Data: Sequential Data

- Bio-sequences
- Discretized music and audio data
- Text

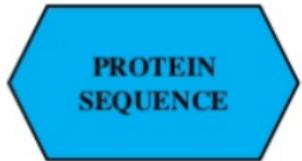
WHAT IS A BIO-SEQUENCE?

DNA, RNA or protein information represented as a series of bases (or amino acids) that appear in bio-molecules. The method by which a bio-sequence is obtained is called *Bio-sequencing*.

GTCCCTGATAAGTCAGTGTCTCC
GAGTCTAGCTTCTGTCCATGCT
GATCATGTCCATGTTCTAGTCAT
GATAGTTGATTCTAGTGTCCCTC



TPPUQWRDCCLKSWCUWMF
ESPWYZWEGHILDDFPTCTWF
CCDTWCUWGHISTDTKKSUN
RGHPPHHLDTWQESRNDCQE



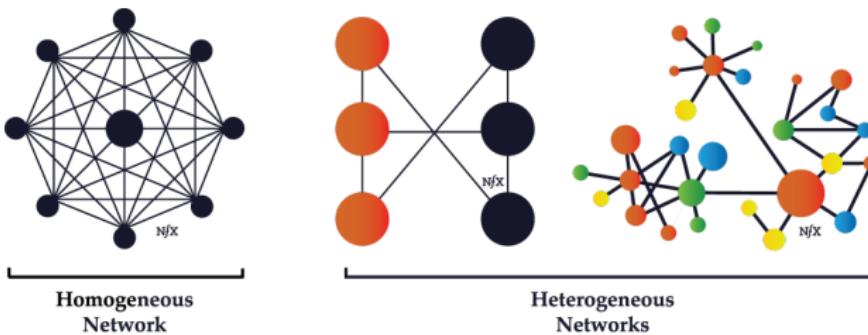
Types of Data: Streams

- Real time data
- Single pass algorithms/online algorithms
- Irreversible decisions
- Small memory algorithms



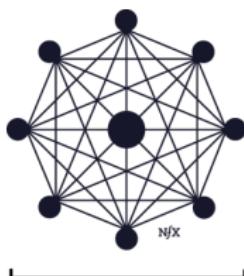
Types of Data: Graphs/Homogeneous Networks

- $G=(V,E)$, data items represented as graphs
- Could have similarity on edges
- Could have weights on vertices, edges or both
- Facebook, webgraph, twitter, co-authorship graphs (bibliometric), citation networks

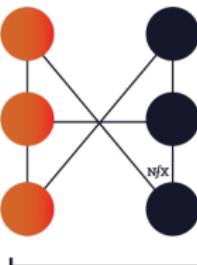


Types of Data: Heterogeneous Networks

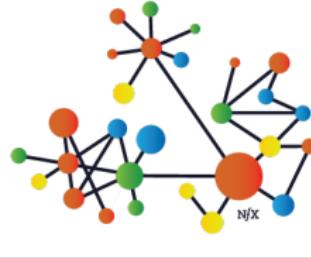
- Nodes represent different entities
- Authors and conferences



Homogeneous
Network



Heterogeneous
Networks



The Analytics Process

■ Business Objective

- Why we are seeking data analytics in the first place?
- How can we reduce production costs without sacrificing quality?
- What are some ways to increase sales opportunities with our current resources?
- Do customers view our brand in a favorable way?

■ Data Collection

- What data is needed and available
- Identify sources of data and relevance of data
- Are there enough instances, are all relevant features there?
- Identify datasets, acquire and retrieve
- Sources RDBMS, .txt, webservices (soup), RSS, tweets
- Experiments, synthetic data generation, Survey

The Analytics Process

■ Data Preparation

- Make the data ready for analytics
- Exploratory Data Analysis Describe, Summarize, Visualize
- Pre-process: Improve data quality, clean data, transformation, standardization, normalization

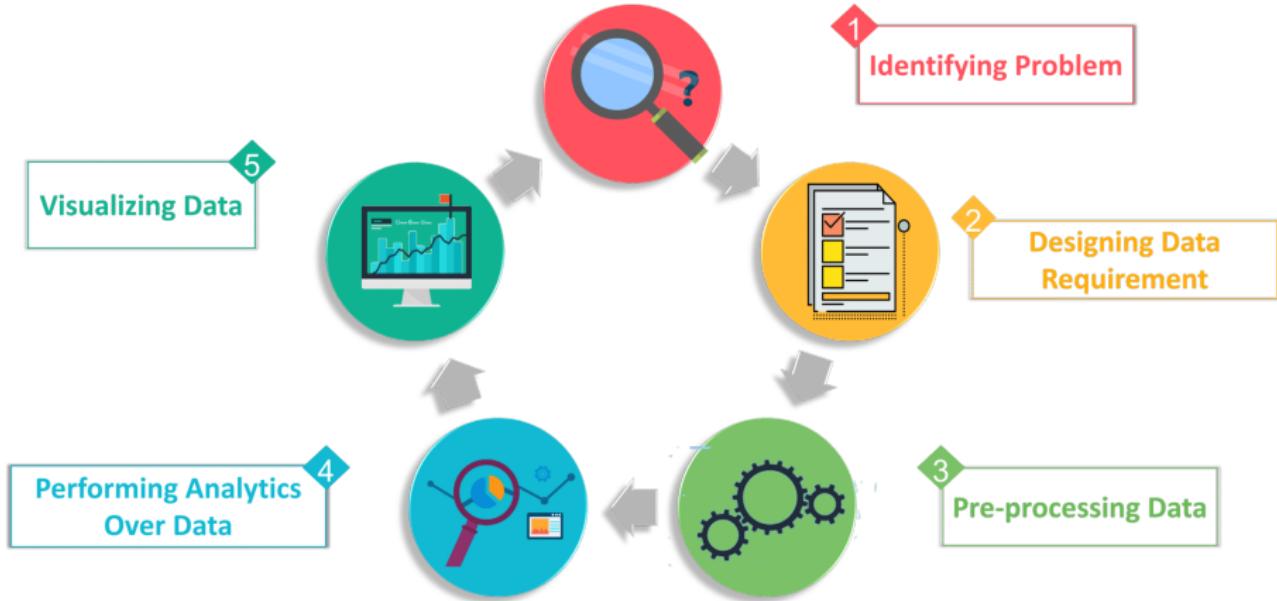
■ Data Analysis

- Apply analytical techniques
- Supervised and unsupervised learning, Graph analytics

■ Report and Deployment

- Communicate results and findings, and apply conclusions to gain benefit

The Analytics Process



The Analytics Process

Machine learning can be combined with other types of analytics to solve a large swath of business problems

