QRELab Quality & Reliability Engineering Laboratory

آزمایشگاه مهندسی کیفیت و پایایی

دانشگاه علم و صنعت ایران

Big Data, Analytics, and the Future of Industrial Engineering's Practice and Research

Amir Gandomi, Ph.D.

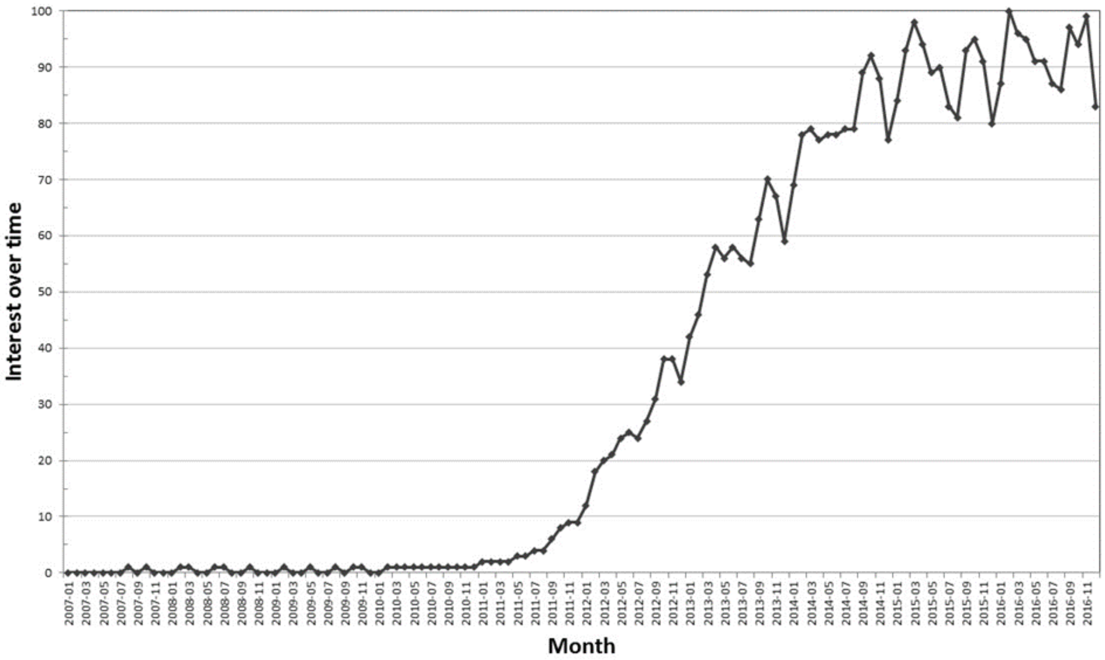
HOFSTRA UNIVERSITY®
FRANK G. ZARB SCHOOL OF BUSINESS

Outline

- What is Big Data?
- Analytics for Unstructured Data
    - Text Analytics
    - Audio Analytics
    - Video Analytics
    - Social Media Analytics
- Big Data Analytics in Marketing
- Key Drivers of Big Data and Analytics
- Analytics Job Market
- Essential Skills of a Modern Data Scientist
- Big Data and Analytics Research

What is Big Data?

- Big Data definitions have evolved rapidly, which has raised some confusion.
  - Many use the term “Big Data” as a buzzword for smarter, more insightful data analysis.
  - Stop using the term “Big Data” (Tom Davenport, 2014)
    - The term “big” is relative.
    - “Big” is only one aspect of what’s distinctive about new forms of data.

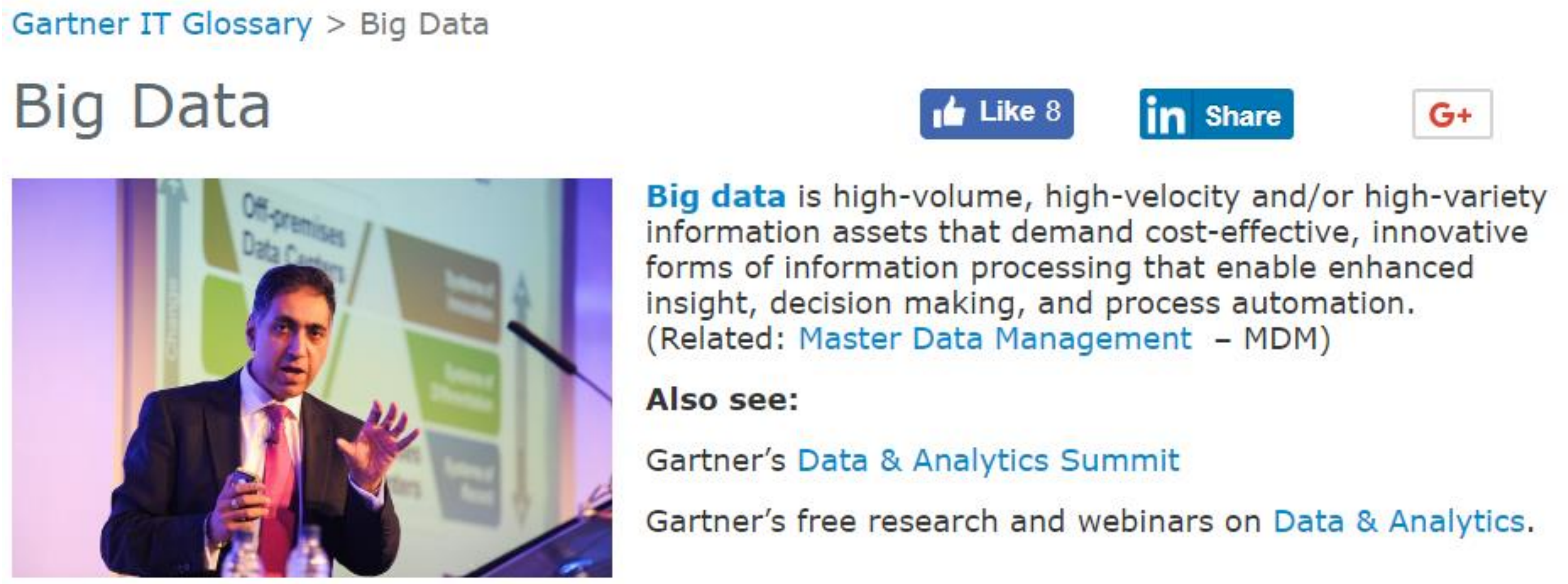


Google Trends for “Big Data” during 2007 to 2016

    - Nobody seems to be comfortable with the opposing term to Big Data, “small data.”
    - Too many people—and vendors in particular—are already using “Big Data” to mean any use of analytics.

What is Big Data? (Continued)

- The **Three V's** have emerged as a common framework to describe Big Data. For example, Gartner, Inc. defines Big Data as follows:



Gartner IT Glossary > Big Data

Big Data

Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation. (Related: Master Data Management – MDM)

Also see:

Gartner's Data & Analytics Summit

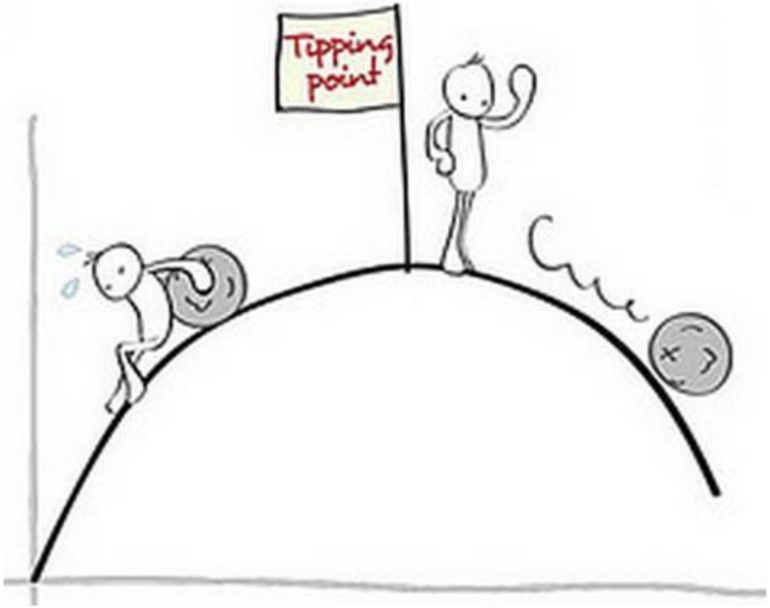
Gartner's free research and webinars on Data & Analytics.

What is Big Data? (Continued)

- ***Volume*** refers to the magnitude of data.
  - Big Data sizes are reported in multiple terabytes and petabytes.
- ***Variety*** refers to the structural heterogeneity in a dataset.
  - *Structured* data, which constitutes only 5% of all existing data, refers to the tabular data found in spreadsheets or relational databases.
  - Text, images, audio, and video are examples of *unstructured* data, which sometimes lack the structural organization required by machines for analysis.
  - Spanning a continuum between fully structured and unstructured data, the format of *semi-structured* data does not conform to strict standards. Extensible Markup Language (XML), a textual language for exchanging data on the Web, is a typical example of semi-structured data.
- **Velocity** refers to the rate at which data is generated and the speed at which it should be analyzed and acted upon.
  - The proliferation of digital devices such as smartphones and sensors has led to an unprecedented rate of data creation and is driving a growing need for real-time analytics and evidence-based planning.

What is Big Data? (Continued)

- **Universal benchmarks** do not exist for volume, variety, and velocity that define Big Data.
  - The defining limits depend upon the size, sector, and location of the firm and these limits evolve over time.
  - Also, these dimensions are not independent of each other.
- A **'three-V tipping point'** therefore exists for every firm beyond which traditional data management and analysis technologies become inadequate for deriving timely intelligence.
- This is where the firm starts dealing with their 'Big Data' and it should evaluate the value created by adopting **Big Data technologies** against the implementation costs.



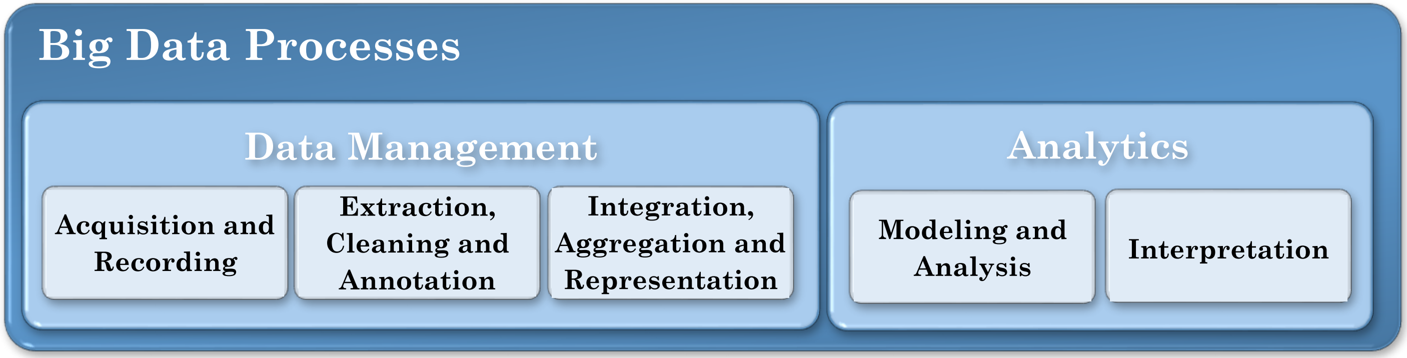
Source: alchemy4thesoul.com

Outline

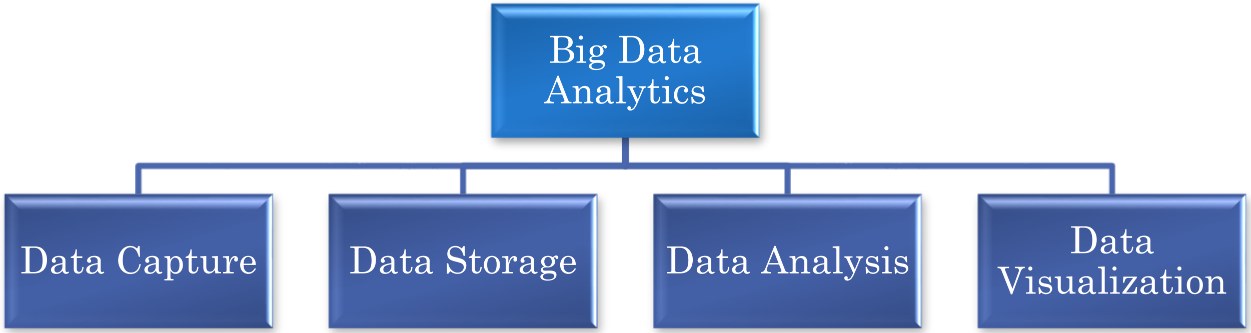
- What is Big Data?
- **Analytics for Unstructured Data**
    - **Text Analytics**
    - **Audio Analytics**
    - **Video Analytics**
    - **Social Media Analytics**
- Big Data Analytics in Marketing
- Key Drivers of Big Data and Analytics
- Analytics Job Market
- Essential Skills of a Modern Data Scientist
- Big Data and Analytics Research

Big Data Analytics

- The overall process of extracting insights from Big Data can be broken down into five stages, shown in the figure below.
  - These five stages form the two main sub-processes: **data management** and **analytics**.

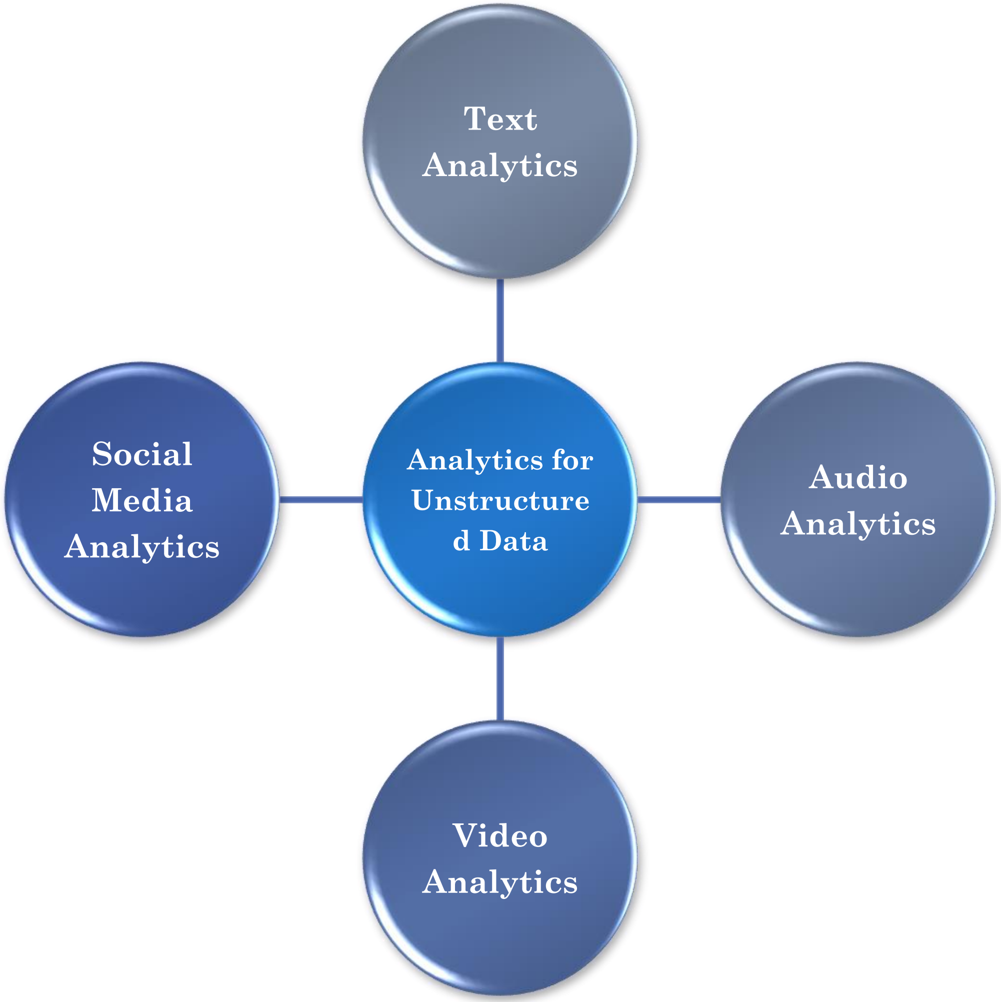

Big Data Processes
Data Management
Acquisition and Recording
Extraction, Cleaning and Annotation
Integration, Aggregation and Representation
Analytics
Modeling and Analysis
Interpretation

- Other frameworks have emerged in recent years, e.g., Portugal et al. (2016):



Emerging Direction in Analytics for Unstructured Data

- In what follows, the emerging analytical techniques for unstructured data are reviewed. Refer to Gandomi and Haider (2015) for a more comprehensive review.



Text Analytics

- **Text analytics** refers to techniques that extract information from textual data.

Sources of data: Social network feeds, emails, blogs, online forums, survey responses, corporate documents, news and call center logs.

Information Extraction (IE)

- To extract structured data from unstructured text.
- Example: derive information such as drug name, dosage and frequency from medical prescriptions.

Text Summarization

- To automatically produce a succinct summary of a single document or multiple documents.
- Example of application areas: scientific and news articles, advertisements, emails and blogs.

Question Answering (QA)

- To provide answers to questions posed in natural language.
- Examples: Apple's Siri and IBM's Watson.

Sentiment Analysis

- To analyze and determine the polarity of opinionated text, which contain people's opinions towards entities such as products, organizations individuals and events.
- Example: Amazon's five-star system.

Text Analytics (Continued)

- The main challenge in text analytics arise from the facts that:
  - Natural language provides the flexibility to convey exactly the same meaning in indefinitely many ways, and
  - Exactly the same statement in a different context may convey completely different meaning.

“This camera really sucks,”
“This vacuum cleaner really sucks”.

- It is extremely difficult to precisely analyze unstructured data in the same way we process structured data. Dialects, jargon, misspellings, acronyms, colloquialism, sarcasm, grammatical complexities, mixing one or more languages in the same text are just some of the fundamental problems unstructured data poses.
- Text analytics is a multidisciplinary field, involving:
  - Data mining,
  - Statistics,
  - Artificial intelligence and machine learning,
  - Computational linguistics,
  - Library and information sciences,
  - Databases.

Audio Analytics

- **Audio analytics** analyze and extract information from unstructured audio data. When applied to human spoken language, audio analytics is also referred to as speech analytics.
- Currently, customer calls centers and healthcare are the primary application areas of audio analytics.

- To efficiently sift through thousands or millions of hours of **recorded calls**. These techniques can be leveraged to improve customer experience, evaluate agent performance, enhance sales turnover rates, gain insight into customer behaviour and identify product or service issues, among many other tasks.

- Significant business value can be drawn using **real-time analytics**. For instance, audio analytics systems can be designed to analyze a live call, formulate cross/up-selling recommendations based on the customer's past and present interactions and provide feedbacks to agents in real-time.

- To Support diagnosis and treatment of certain medical conditions that affect the patient's communication patterns (e.g., depression, schizophrenia and cancer).

- Help to analyze an infant's cry, which carries various information about the infant's health and emotional status.

Audio Analytics (Continued)

- Speech analytics follows two common technological approaches: the **transcript-based approach** (widely known as large-vocabulary continuous speech recognition, LVCSR) and the **phonetic-based approach**. These are explained below.

LVCSR systems follow a two-phase process: **indexing and searching**. In the first phase, they attempt to transcribe the speech content of the audio. This is performed using automatic speech recognition (ASR) algorithms that match sounds to words. The words are identified based on a predefined dictionary. In the second phase, standard text-based methods are used to find the search term in the index file.

Phonetic-based systems work with sounds or ***phonemes***. They also consist of two phases: phonetic indexing and searching. In the first phase, the system translates the input speech into a sequence of phonemes. In the second phase, the system searches the output of the first phase for the phonetic representation of the search terms.

Video Analytics

- **Video analytics**, also known as video content analysis (VCA), involves a variety of techniques to monitor, analyze, and extract meaningful information from video streams.
- The increasing prevalence of closed-circuit television (CCTV) cameras and the booming popularity of video-sharing websites like YouTube are two factors that have contributed to the growth of computerized video analysis.
- A key challenge is the sheer size of video data. To put this into perspective:
  - One second of a high-definition video = over 2,000 pages of text,
  - Every minute, 100 hours of video are uploaded to YouTube.
- The primary application of video analytics has been in the automated security and surveillance systems:
  - Detecting breaches of restricted zones,
  - Identifying objects removed or left unattended,
  - Detecting loitering in a specific area,
  - Recognizing suspicious activities and detecting camera tampering.

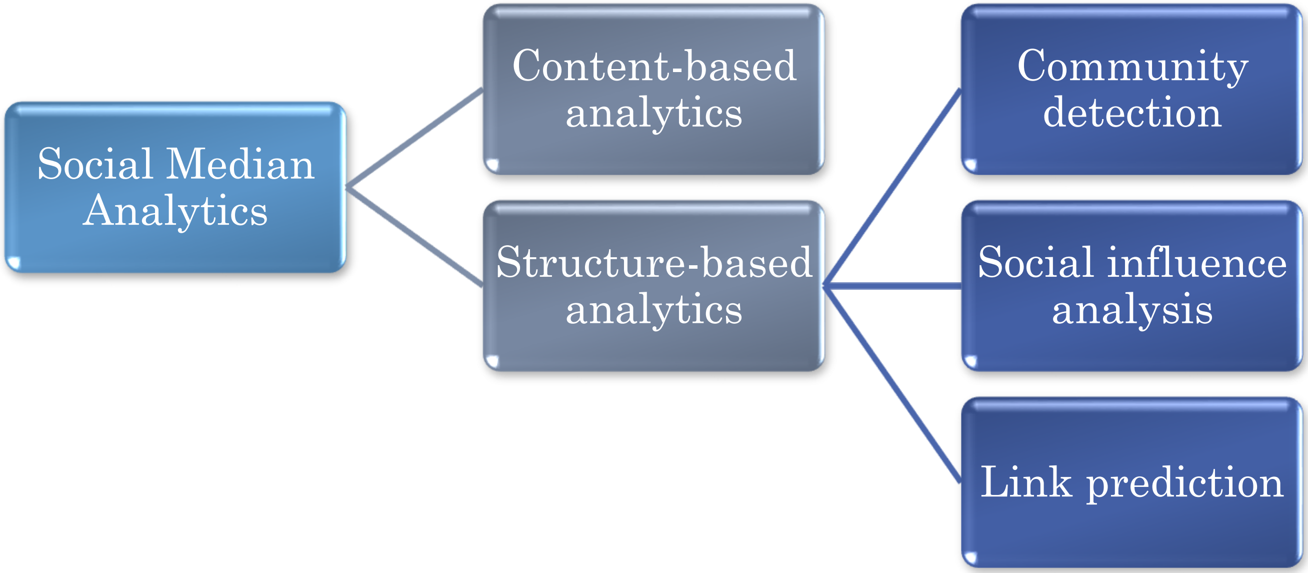
Video Analytics (Continued)

- The data generated through CCTV cameras in retail outlets can be exploited to extract business insights:
  - To collect demographic information of the customer population.
  - To count the number of customers, measure the time they stay in the store, detect their movement patterns, measure their dwell time in different areas and monitor queues in real time.
    - This information can be used to drive **various decisions** such as product placement, price and assortment optimization, promotion design, cross-selling, layout optimization and staffing.
  - To personally identify a customer and use their shopping profile for instant and personalized marketing. Such applications, which are commonly referred to **soft surveillance**, raise serious privacy concerns that should be addressed prior to adoption.
  - To study of buying behavior of groups. Among family members who shop together, only one interacts with the store at the cash register, causing the traditional systems to miss data on buying patterns of other members.

Social Media Analytics

- **Social media analytics** refers to the computational models used to analyze structured and unstructured data from social media channels:
  - **Social networks:** Facebook and LinkedIn,
  - **Blogs:** e.g., Blogger and WordPress,
  - **Microblogs:** Twitter and Tumblr,
  - **Social news:** Digg and Reddit,
  - **Social bookmarking:** Delicious and StumbleUpon,
  - **Media sharing:** Instagram and YouTube,
  - **Wikis:** Wikipedia and Wikihow,
  - **Question-and-answer sites:** Yahoo! Answers and Ask.com,
  - **Review sites:** Yelp, TripAdvisor.
- The research on social media analytics spans across several disciplines including psychology, sociology, anthropology, computer science, mathematics, physics and economics.

Social Media Analytics (Continued)



- **Content-based analytics** is focused on the data that users post on social media platforms. Text, audio and video analytics can be applied to derive insight from such data.
- **Structure-based analytics** is concerned with synthesizing the structural attributes of a social network and extracting intelligence from the relationships among the participating entities.

Social Media Analytics (Continued)

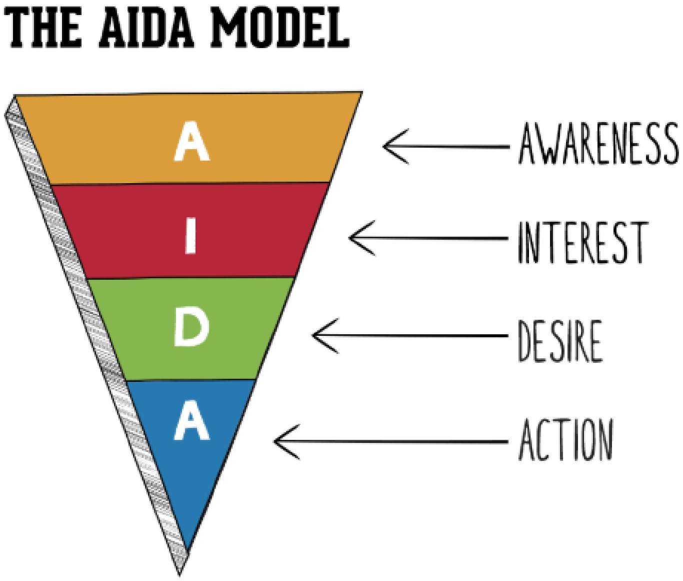
- **Community detection** extracts implicit communities within a network.
  - It helps to summarize such huge networks, which facilitates the process of uncovering the existing behavioral patterns and predicting the emergent properties of the network
  - Application: Developing more effective product recommendation systems.
- **Social influence analysis** is concerned with modelling and evaluating the influence of actors and connections in a social network.
  - Application: Viral marketing to efficiently enhance brand awareness and adoption.
- **Link prediction** specifically addresses the problem of predicting future linkages between the existing nodes in the underlying network.
  - In security, link prediction helps to uncover potential collaborations in terrorist or criminal networks.
  - In science, link prediction can be applied to discover co-authorship opportunities
  - In the context of online social media, the primary application of link prediction is in the development of recommendation systems such as Facebook's "People You May Know", YouTube's "Recommended for You" and Netflix's and Amazon's recommender engines.

Outline

- What is Big Data?
- Analytics for Unstructured Data
  - Text Analytics
  - Audio Analytics
  - Video Analytics
  - Social Media Analytics
- **Big Data Analytics in Marketing**
- Key Drivers of Big Data and Analytics
- Analytics Job Market
- Essential Skills of a Modern Data Scientist
- Big Data and Analytics Research

Big Data Analytics in Marketing

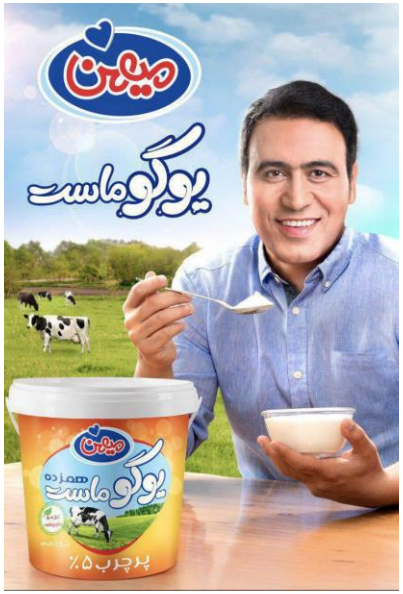
- As a data-rich field, **marketing** has emerged as one of the primary application areas of Big Data analytics.
- To better understand the value of Big Data analytics in marketing, consider the **AIDA model**:
  - The AIDA Model identifies **cognitive stages** an individual goes through during the buying process for a product or service.
    - **Awareness:** The consumer becomes aware of a category, product or brand (usually through advertising)
    - **Interest:** The consumer becomes interested by learning about brand benefits & how the brand fits with lifestyle



Source: smartinsights.com

    - **Desire:** The consumer develops a favorable disposition towards the brand.
    - **Action:** The consumer forms a purchase intention, shops around, engages in trial or makes a purchase.

Big Data Analytics in Marketing (Continued)



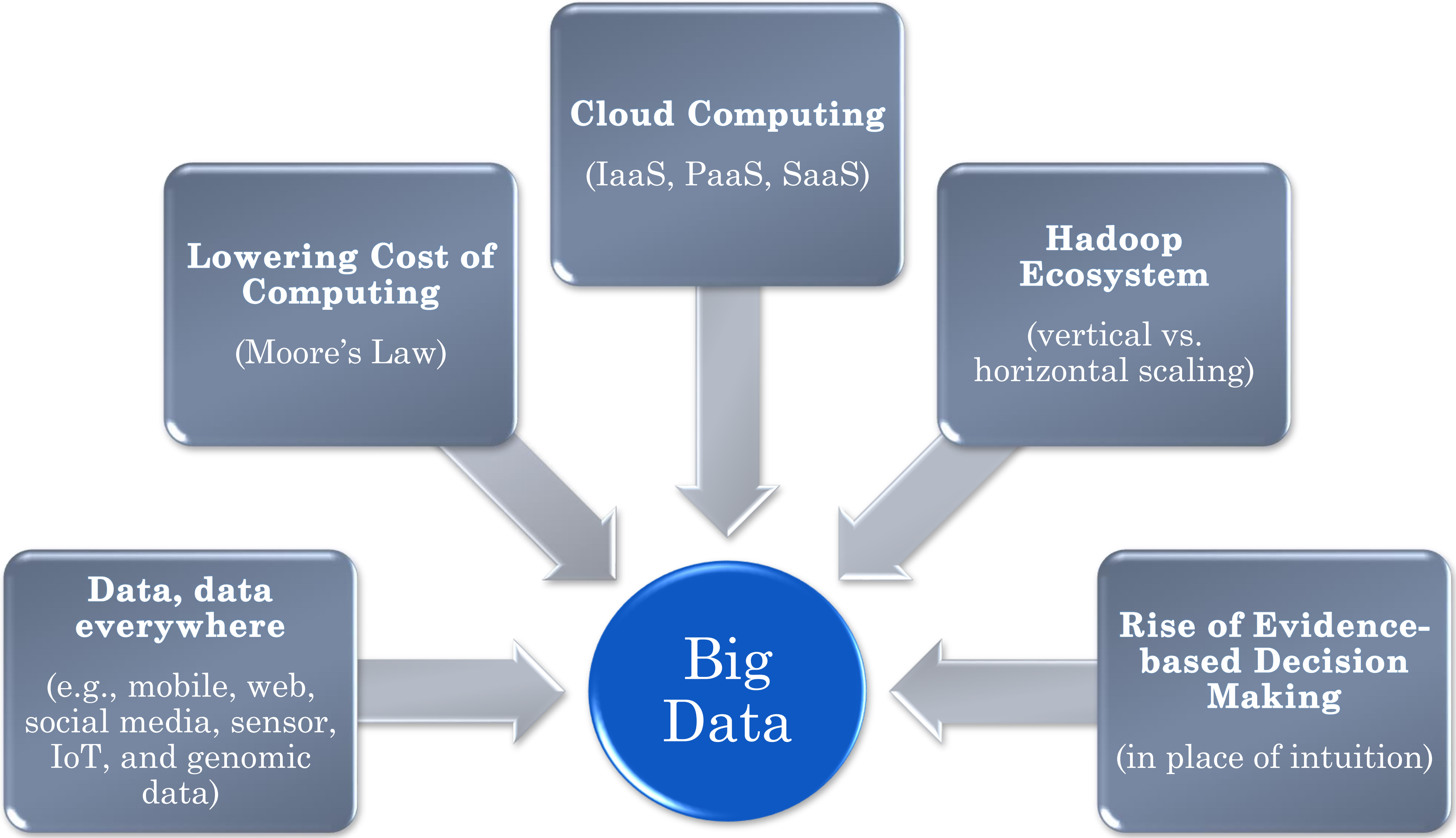
Big Data Analytics in Marketing (Continued)

- Previously, only the Action stage resulted in hard data that could be recorded and analyzed as secondary data.
- Data pertaining to the other stages (A, I and D) had to be obtained through primary means for limited samples in target segments.
- With the advent of Big Data, the path to purchase, and where it is “stopped” is now better understood.
- This is enabled by a data from a variety of sources including:
  - **Social media**, **search**, and **clickstream** data in online settings,
  - **Geolocation**, **RFID**, and **video** data in the brick-and-mortar retail.

Outline

- What is Big Data?
- Analytics for Unstructured Data
  - Text Analytics
  - Audio Analytics
  - Video Analytics
  - Social Media Analytics
- Big Data Analytics in Marketing
- **Key Drivers of Big Data and Analytics**
- Analytics Job Market
- Essential Skills of a Modern Data Scientist
- Big Data and Analytics Research

Key Drivers of Big Data



Cloud Computing

(IaaS, PaaS, SaaS)

Lowering Cost of Computing

(Moore's Law)

Hadoop Ecosystem

(vertical vs. horizontal scaling)

Data, data everywhere

(e.g., mobile, web, social media, sensor, IoT, and genomic data)

Big Data

Rise of Evidence-based Decision Making

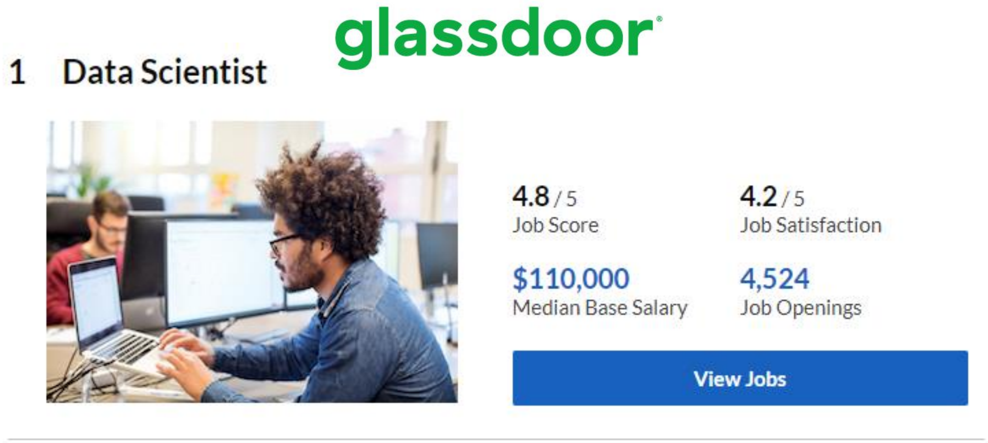
(in place of intuition)

© Amir Gandomi, Frank G. Zarb School of Business

Outline

- What is Big Data?
- Analytics for Unstructured Data
    - Text Analytics
    - Audio Analytics
    - Video Analytics
    - Social Media Analytics
- Big Data Analytics in Marketing
- Key Drivers of Big Data and Analytics
- **Analytics Job Market**
- Essential Skills of a Modern Data Scientist
- Big Data and Analytics Research

Analytics Job Market



glassdoor

1 Data Scientist

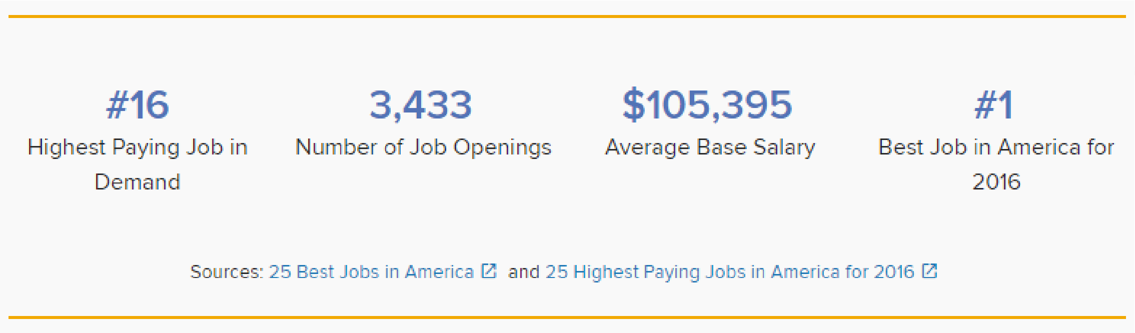
4.8 / 5
Job Score

4.2 / 5
Job Satisfaction

$110,000
Median Base Salary

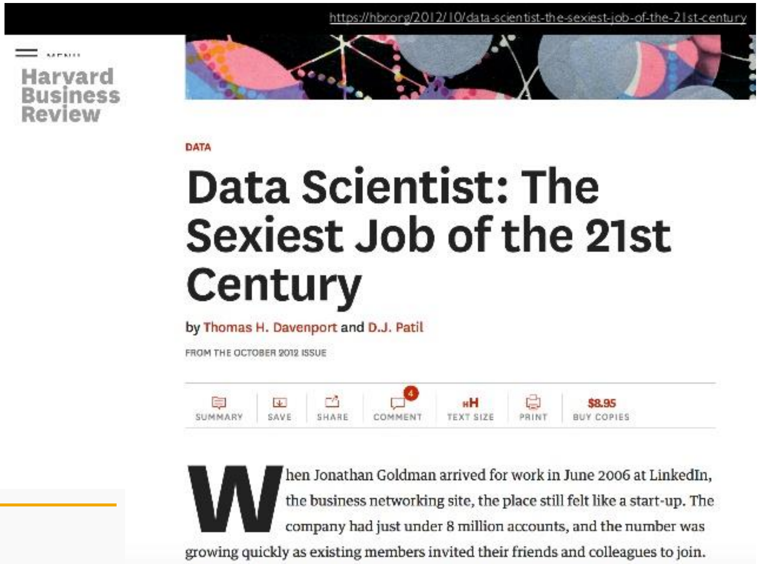
4,524
Job Openings

View Jobs



#16	3,433	$105,395	#1
Highest Paying Job in Demand	Number of Job Openings	Average Base Salary	Best Job in America for 2016

Sources: 25 Best Jobs in America and 25 Highest Paying Jobs in America for 2016



https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

Harvard Business Review

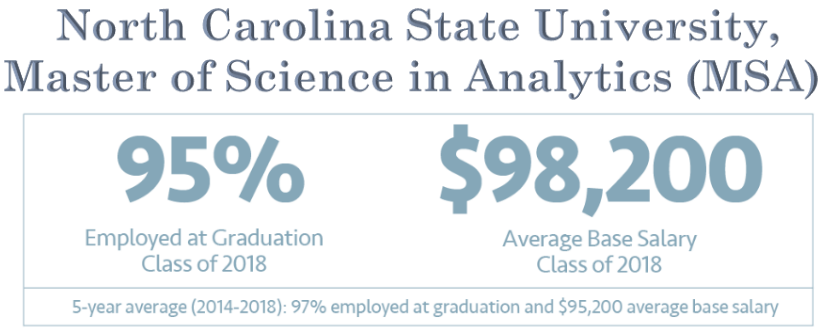
DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join.



North Carolina State University, Master of Science in Analytics (MSA)

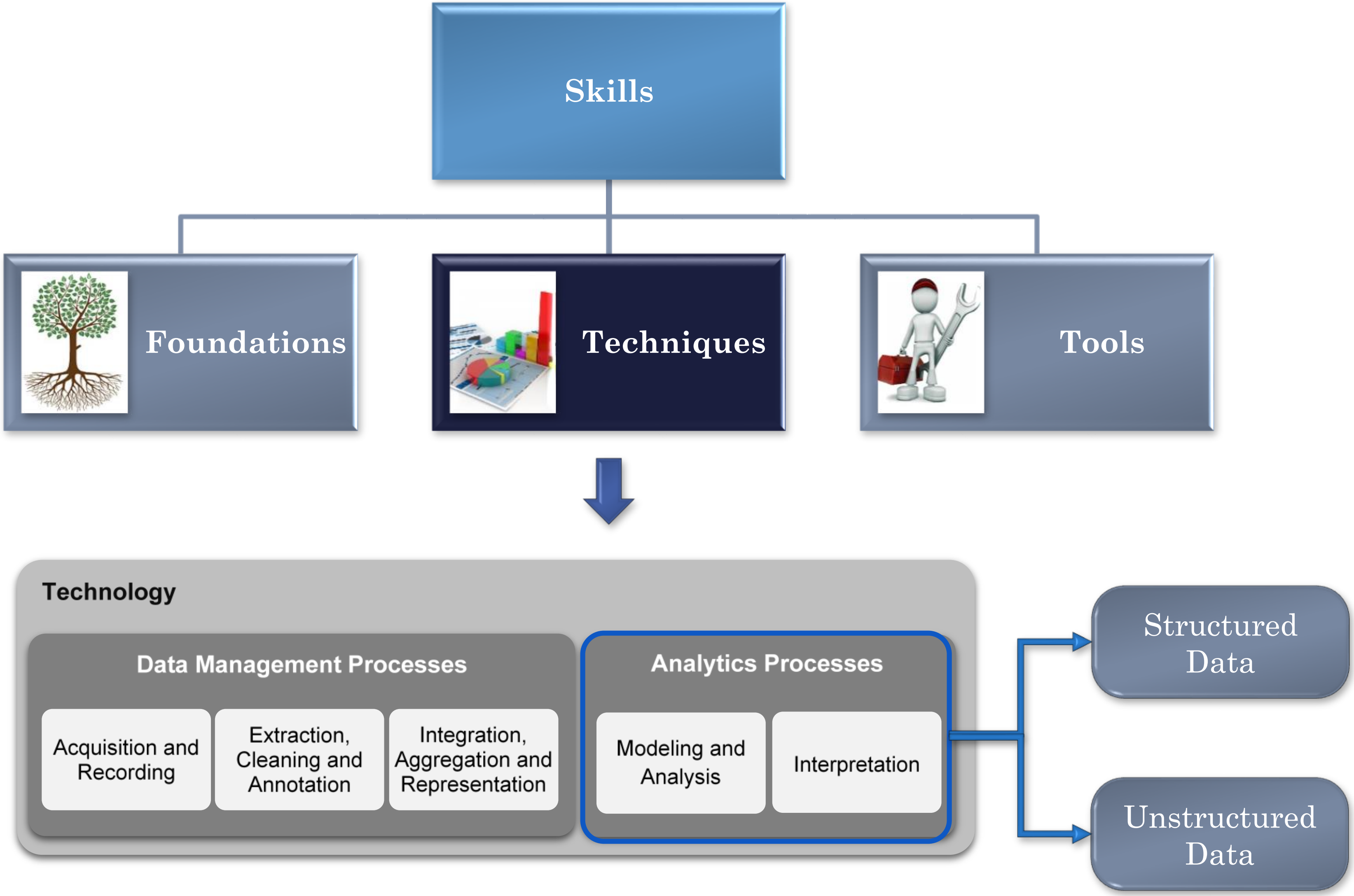
95%	$98,200
Employed at Graduation Class of 2018	Average Base Salary Class of 2018

5-year average (2014-2018): 97% employed at graduation and $95,200 average base salary

Outline

- What is Big Data?
- Analytics for Unstructured Data
  - Text Analytics
  - Audio Analytics
  - Video Analytics
  - Social Media Analytics
- Big Data Analytics in Marketing
- Key Drivers of Big Data and Analytics
- Analytics Job Market
- **Essential Skills of a Modern Data Scientist**
- Big Data and Analytics Research

Essential Skills of a Modern Data Scientist



Skills

Foundations

Techniques

Tools

Technology

Data Management Processes

Acquisition and Recording

Extraction, Cleaning and Annotation

Integration, Aggregation and Representation

Analytics Processes

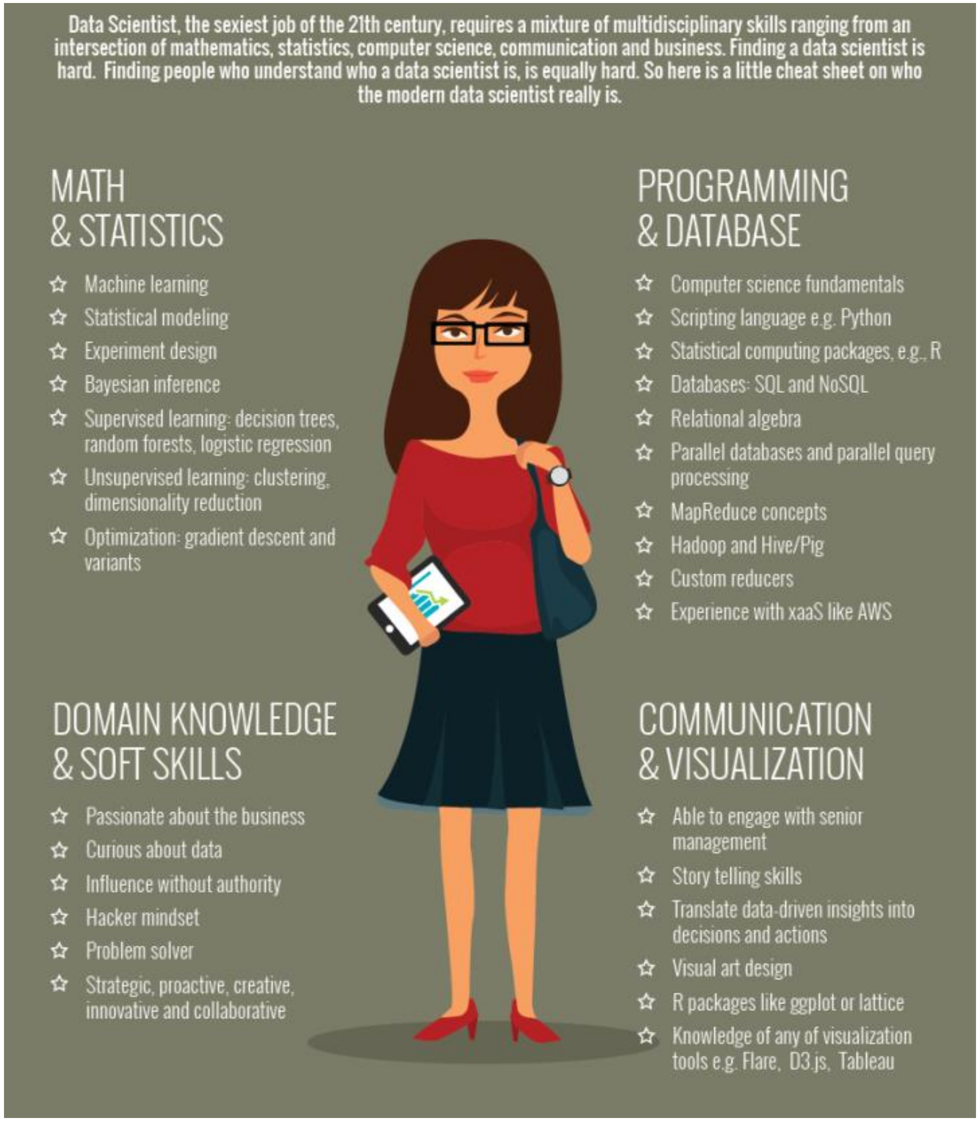
Modeling and Analysis

Interpretation

Structured Data

Unstructured Data

Essential Skills of a Modern Data Scientist – Foundations



Data Scientist, the sexiest job of the 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 packages, 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

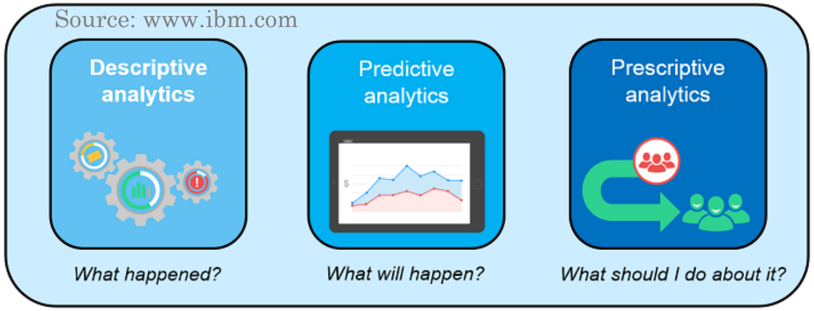
- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

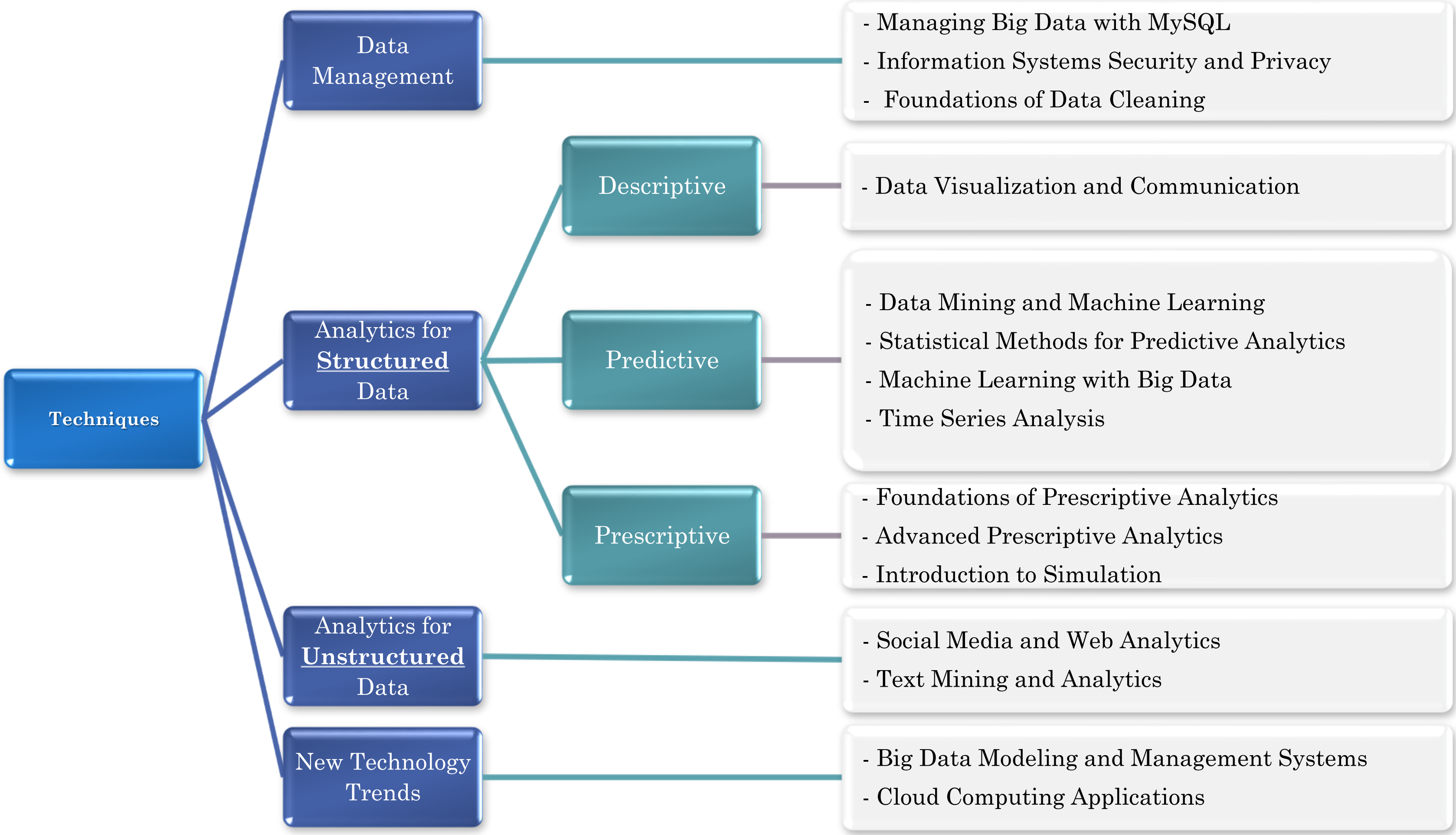
- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

Source: schoolofdatascience.amsterdam

Essential Skills of a Modern Data Scientist – Techniques

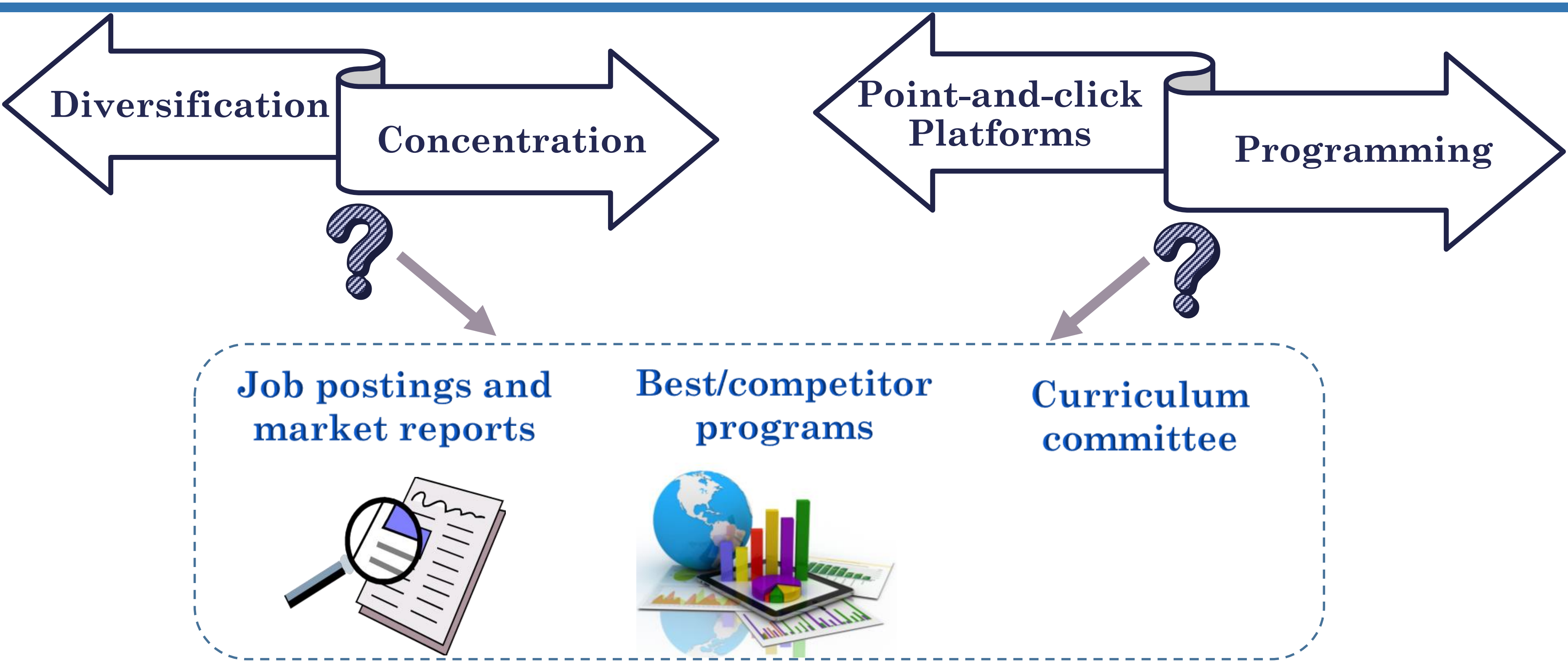


Source: www.ibm.com



- Techniques
  - Data Management
    - Managing Big Data with MySQL
    - Information Systems Security and Privacy
    - Foundations of Data Cleaning
  - Analytics for **Structured** Data
    - Descriptive
      - Data Visualization and Communication
    - Predictive
      - Data Mining and Machine Learning
      - Statistical Methods for Predictive Analytics
      - Machine Learning with Big Data
      - Time Series Analysis
    - Prescriptive
      - Foundations of Prescriptive Analytics
      - Advanced Prescriptive Analytics
      - Introduction to Simulation
  - Analytics for **Unstructured** Data
    - Social Media and Web Analytics
    - Text Mining and Analytics
  - New Technology Trends
    - Big Data Modeling and Management Systems
    - Cloud Computing Applications

Essential Skills of a Modern Data Scientist – Tools



Diversification

Concentration

Point-and-click Platforms

Programming

Job postings and market reports

Best/competitor programs

Curriculum committee

sas SAP Java TERADATA salesforce IBM Watson

ORACLE R tableau MySQL APACHE Spark

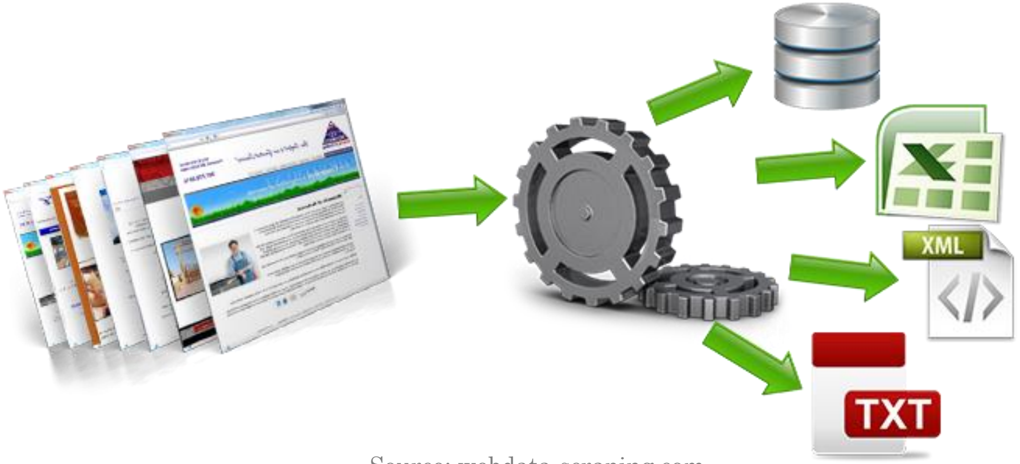
Microsoft IBM python rapidminer hadoop

Outline

- What is Big Data?
- Analytics for Unstructured Data
    - Text Analytics
    - Audio Analytics
    - Video Analytics
    - Social Media Analytics
- Big Data Analytics in Marketing
- Key Drivers of Big Data and Analytics
- Analytics Job Market
- Essential Skills of a Modern Data Scientist
- **Big Data and Analytics Research**

Big Data and Analytics Research

- New research insight often arises either from **new data**, **new methods**, or the **combination of the two** (Bradlow et al., 2017).
- **Web scraping** can be used to extract new data from the web.



Source: webdata-scraping.com

in

RATE MY PROFESSORS

GROUPON

- **Open Data** initiatives is another source of new data.

DATA.GOV

THE WORLD BANK

CANADIAN OPEN DATA EXPERIENCE

kaggle

با تشکر