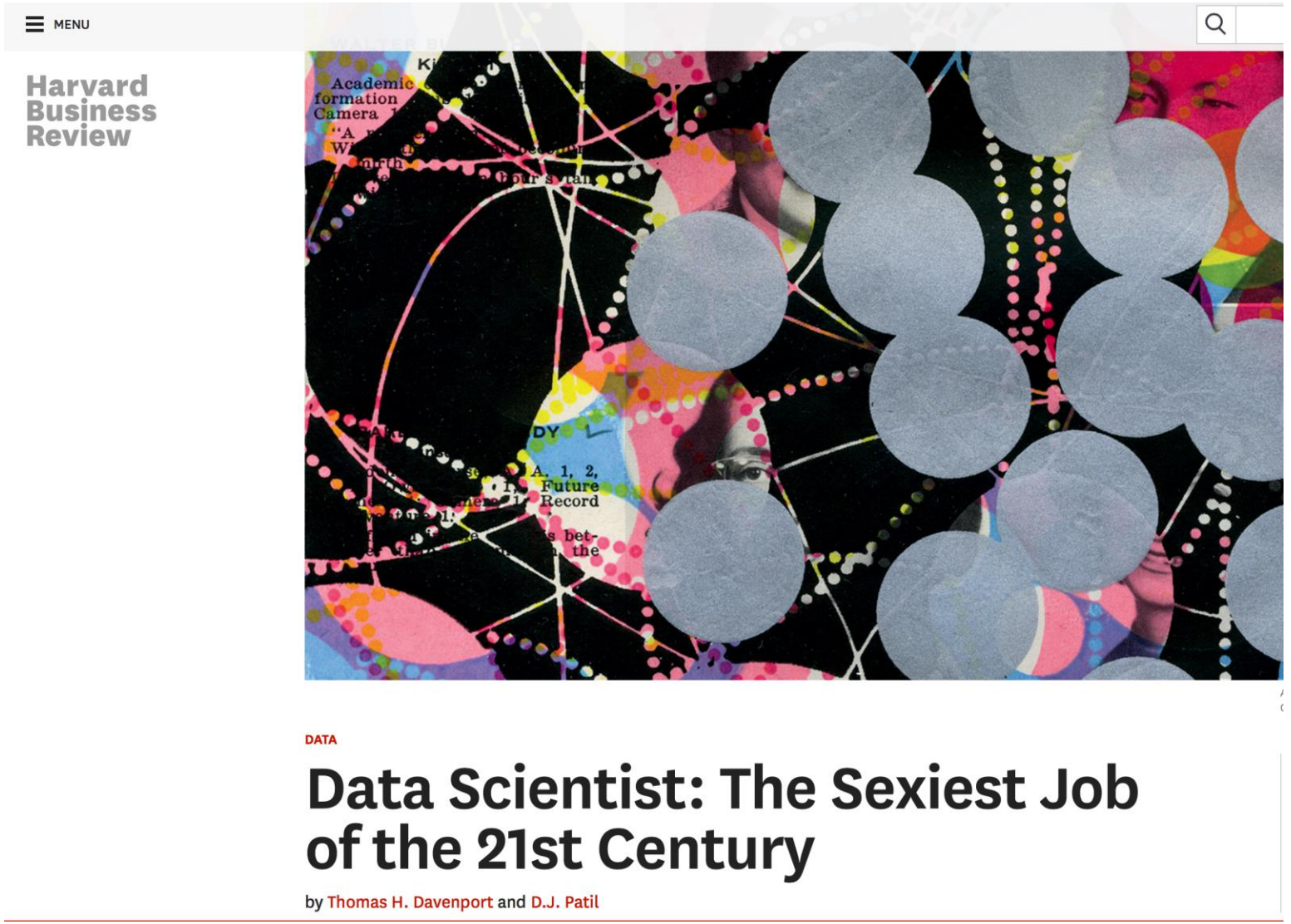
Big Data – An Introduction

Instructor, Nero Chan Zhen Yu

Forward School

Sexiest job in 21st Century

Screenshot taken from HBR article: https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century



Forward School

Data is the new oil

The Economist | Topics | Current edition | More

Screenshot taken from The Economist article https://www.economist.com/news/leaders/21721656-data-economy-demands-new-approach-antitrust-rules-worlds-most-valuable-resource

Regulating the internet giants

The world's most valuable resource is no longer oil, but data

The data economy demands a new approach to antitrust rules

Forward School

Hype or ~~$$$~~Reality?

- Everyone is talking about
- Everyone plans to learn something about it
- Everyone knows little what/how to do

Forward School

What is Analytics?

- Analytics is a broad term that encompasses ***the processes, technologies, frameworks and algorithms*** to ***extract meaningful insights*** from **data**.
- Raw data - has no meaning by itself until
  - It is contextualized and processed into useful information
- Typical processes - Extracting and creating information from raw data by
  - Filtering, processing, categorizing, condensing and contextualizing the data
- This information – when organized and structured
  - Can infer knowledge about the system, its users, its environment and its operations and progress towards its objectives.
  - Thus making the system smarter and more efficient.

Forward School

What is Analytics?

- The choice of technologies, algorithms, frameworks for analytics
  - Driven by the analytics goal of the application
- Example goals
  - Predict whether a transaction is fraudulent
  - Predict whether a tumor is benign or malignant
  - Find patterns of coldest day in the year
  - Find most visited pages on a particular website
  - Find similar patients in an electronic health record systems
  - Find correlation between news items and stock prices

Forward School

Four types of analytics

- Descriptive analytics
- Diagnostic analytics
- Predictive analytics
- Prescriptive analytics

Forward School

Descriptive Analytics

- Descriptive Analytics comprises analyzing past data
  - To present it in a summarized form
  - For easier interpretation
- Descriptive Analytics aims to answer – ***What has happened?***
- A major portion of analytics done today is descriptive analytics
  - Done through the use of statistics
  - Is useful to summarize the data

Forward School

Diagnostic Analytics

- Diagnostic analytics comprises analysis of past data to diagnose the reasons
  - As to why certain events happened
- Diagnostic analytics aims to answer – ***Why did it happened?***
- It does more than just summarizing data through statistics (descriptive analytics)
  - Aims to provide more insights into why certain incident has occurred based on the patterns identified
  - E.g. patterns in sensor data may explain why a fault has occurred.

Forward School

Predictive Analytics

- Predictive analytics comprises predicting the occurrence of an event
  - Likely outcome of an event
  - Forecasting the future values using predictive models
- Predictive analytics aims to answer – ***What is likely to happen?***
- It is done using predictive models which are trained by existing data
  - Example: when a fault will occur in a machine, when a tumor is benign/malignant, predicting the occurrence of natural emergency

Forward School

Predictive Analytics

- The accuracy of the prediction
  - Depends on the quality and the volume of the existing data available during the training phase
  - so that all patterns and trends in the existing data can be learned accurately.
- Validation of the model – necessary before a model is used for prediction
- Typical approach – divide the existing data into training and test data sets
  - Say 75% training vs 25% testing
  - Leave-one-out Cross validation (LOOCV), n-fold Cross Validation (CV) etc.

Forward School

Prescriptive Analytics

- Prescriptive analytics uses multiple prediction models to predict various outcomes and the best course of action for each outcome
- Prescriptive analytics aims to answer – ***What can we do to make it happen?***
- Prescriptive analytics can predict the possible outcomes based on the current choice of actions.

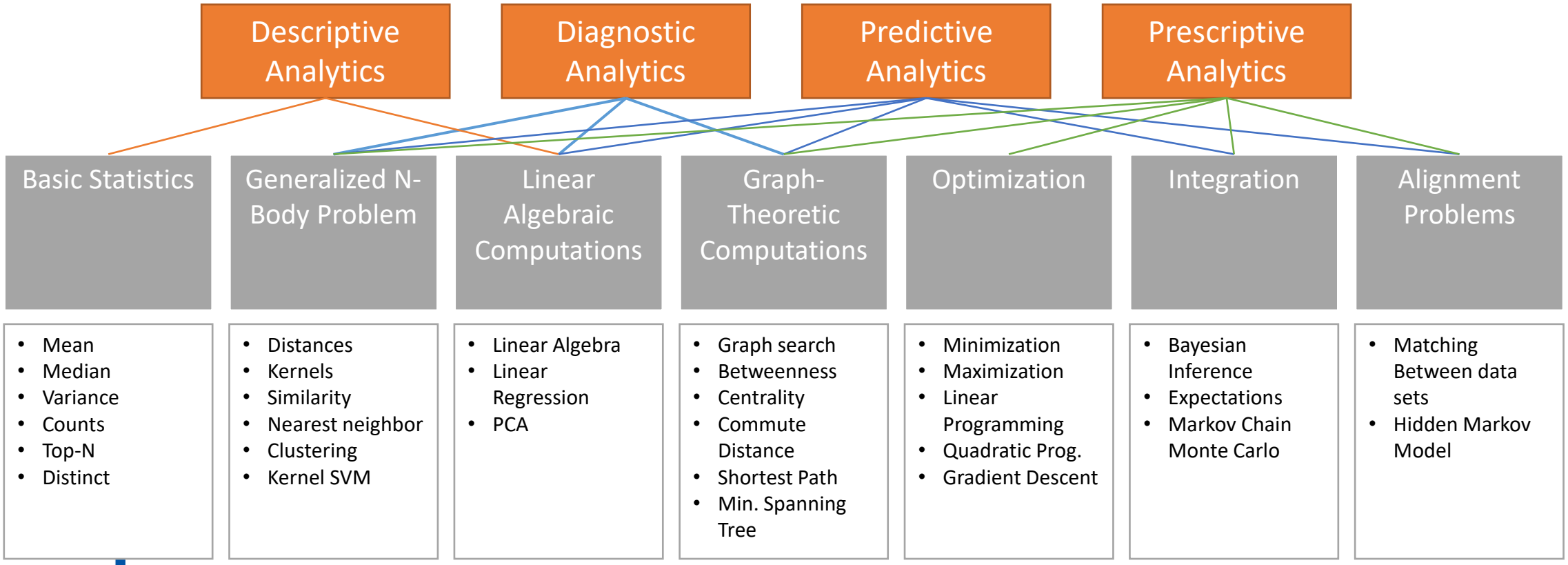
Forward School

Prescriptive Analytics

- We can consider prescriptive analytics as a type of analytics that uses different prediction models for different inputs
  - It prescribes actions or the best options to follow from the available options
- Example: Prescriptive analytics can be used
  - To prescribe the best medicine for treatment of a patient based on the outcomes of various medicines for similar patients.
  - To suggest the best mobile data plan for a customer based on the customer's browsing patterns

Forward School

Mapping between types of Analytics (Figure 1.1)



Forward School

BIG Data – ~~WTF?~~

- Can be defined as collections of datasets whose volume, velocity and variety (3Vs or V^3) is so large that it is
  - Difficult to store, manage, process and analyze the data using conventional/traditional databases and data processing tools.
- In recent years
  - An exponential growth in the both structured and unstructured data generated by information technology, industrial, healthcare, Internet of Things (IoT) and other systems.

Forward School

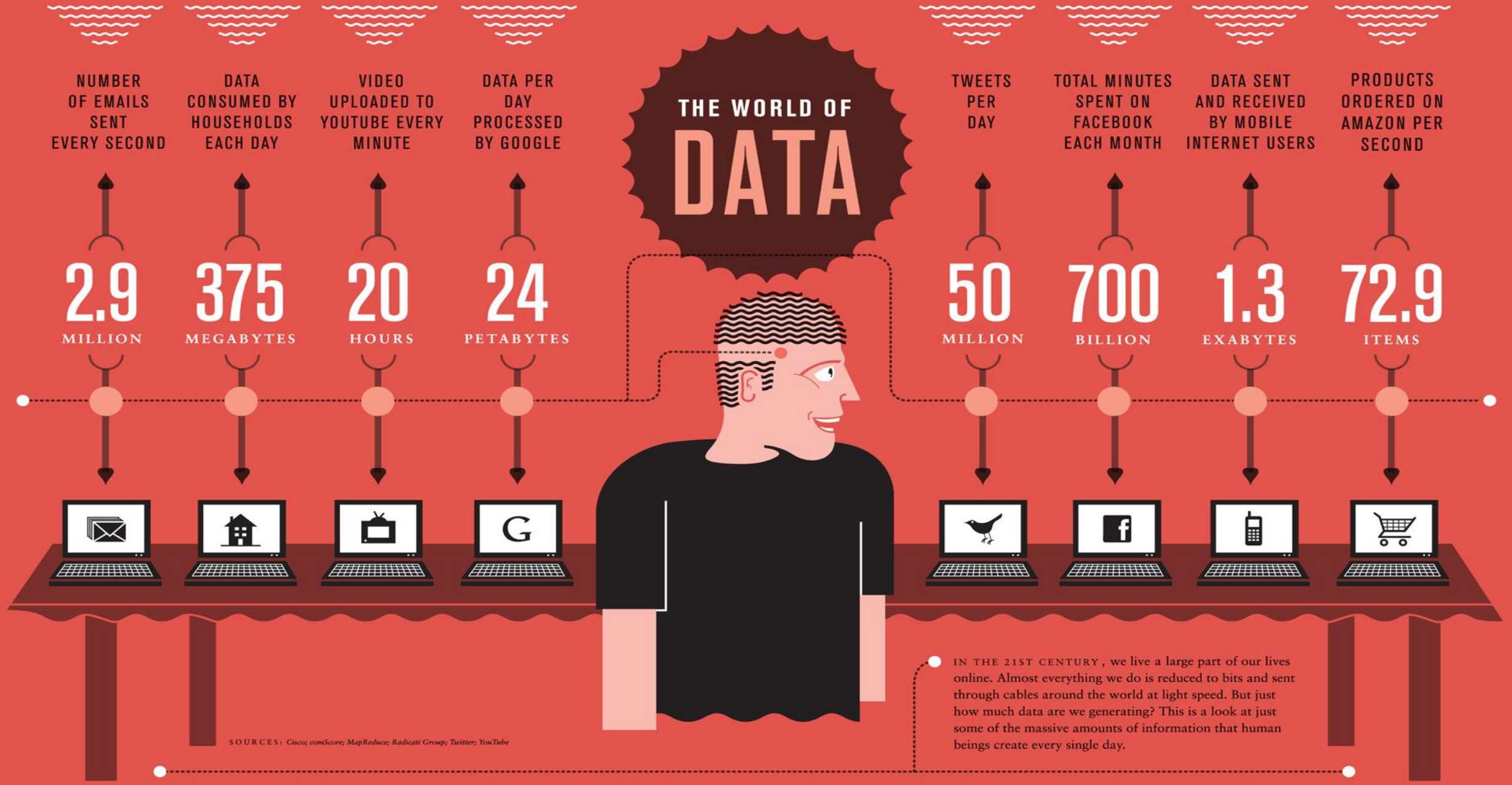
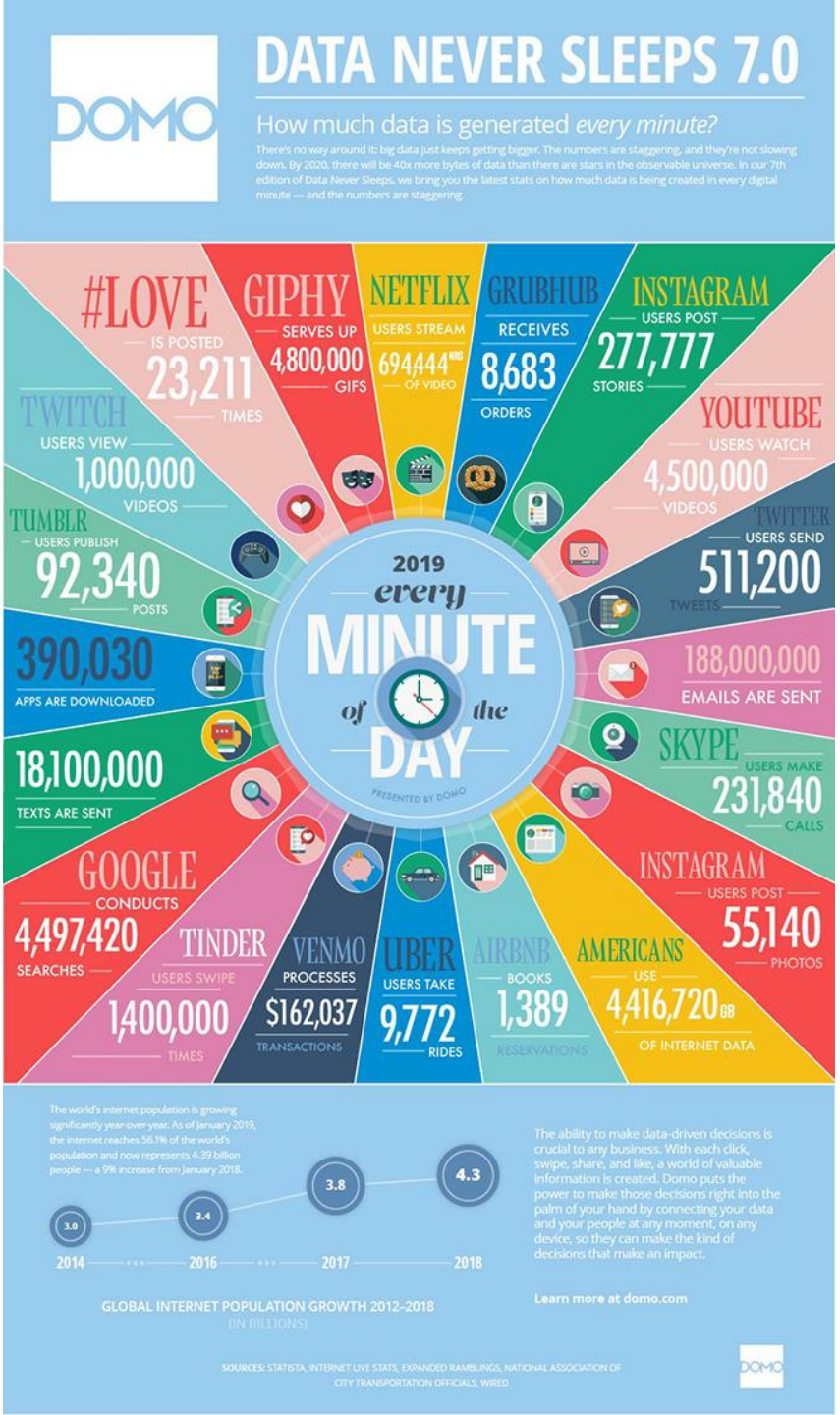


Image taken from https://greenbookblog.org/2013/05/16/are-you-burning-away-your-data-fuel/



Huge amount of data every minute

- But are all useful?
- How to make them useful?
- Noise vs Useful data…

Image taken from Domo: Data never sleeps 7.0 - https://web-assets.domo.com/blog/wp-content/uploads/2019/07/data-never-sleeps-7-896kb.jpg

Power, potential or just marketing?

- Can Big Data really solve ~~all~~ problems on earth?
- How can Big data solve (all) problems on earth?

- It does have the potential to solve many problems
  - If we can leverage the power of data and the tools that allow suitable analytics to be carried out
  - Then we can have the smarter/more intelligent applications

Forward School

Big data analytics (BDA)

- BDA deals with collection, storage, processing and analysis of this massive-scale data
- Specialized tools and frameworks are required for BDA when
  - The volume of data involved is so large that it is difficult to store, process and analyze data on a single machine
  - The velocity of the data is very high and the data needs to be analyzed in real-time
  - The variety of data involved may be structured, unstructured or semi-structured, and the data is collected from multiple (non-similar) data sources
  - Various types of analytics need to be performed to extract value from the data (including all four types of analytics)

Forward School

Big data analytics (BDA)

- BDA tools and frameworks
  - have distributed and parallel processing architectures
  - Can leverage the storage and computational resources of a large cluster of machines
- Typical steps in BDA
  - Data cleansing
  - Data munging(wrangling)
  - Data processing
  - Data visualization

Forward School

BDA Life cycle

- Data ingestion
  - Collect data from multiple data sources
  - Specialized tools and framework are used to ingest this data into the BDA backend
- Data Storage
  - BDA specialized data storage solutions are used
  - Such as distributed filesystems and non-relational databases
  - These storage solutions are designed to scale
- Data analysis
  - Specialized framework is selected based on types of analysis and analysis requirements (batch or real-time)

What are some examples of big data analysis you can name?

- Data generated by social networks including text, images, audio and video data
- Click-stream data generated by web applications such as e-Commerce to analyze user behavior
- Machine sensor data collected from sensors embedded in industrial and energy systems (think Industry 4.0)
- Healthcare data collected in electronic health record (EHR) systems
- Logs generated by web applications
- Stock markets data
- Transactional data generated by banking and financial applications

Forward School

Characteristics of Big Data

Forward School

Typical characteristics of Big Data

- ***The 5Vs of Big Data***
  - Volume
  - Velocity
  - Variety
  - Veracity
  - Value

- At the beginning of (the Hype of) Big data, many only spoke about V^3
  - Volume, Velocity and Variety

Forward School

Volume

- Volume refers to the sheer amount data that no longer fit on a single machine
  - Hence the need of specialized tools and frameworks
- Reasons ~~(culprits)~~ are
  - Lowering costs and availability of sensors and data sources
  - Lowering costs of data storage
- No fixed threshold how big should the volume be to qualify as big data
  - Typically the term big data is used for massive scale data that is difficult to store, manage and process using traditional databases and data processing architectures

Forward School

Velocity

- Velocity refers to how fast the data is generated
- Data can arrive at a data storage at very high velocities
  - Examples: social media data, sensor data
- High velocity of data results in the volume of data accumulated to become very large in short span of time
- Furthermore, some applications has strict deadlines for data to be analyzed
  - Hence, the data needs to be analyzed in real-time
- Specialized tools are required to ingest such high velocity data into the big data infrastructure and analyze the data in real-time

Forward School

Variety

- Variety refers to the forms of the data
- Big data (has the expectation that) comes in different forms such as structured, unstructured and semi-structured
  - Example – text data, audio, image, video, sensor data
- BDA systems needs to be flexible enough to handle such variety of data

Forward School

Veracity

- Veracity refers to how accurate is the data
- To extract value from the data, data needs to be cleaned to remove noise
- Benefits of big data can be obtained from data-driven applications
  - Only when the data is **meaningful** and **accurate**
- If there is any incorrect data, noise, or faulty data
  - Filtering and cleaning are absolutely important

Forward School

Value

- Value of data
  - Refers to the usefulness of data for the intended purpose
- End goal of BDA – to extract value from the data
- Not all data is useful, let alone brings value
- Value of the data is also related to the veracity or accuracy of data
- For some applications, value also depends on how fast one can process the data
  - Value is lost when insights are found too late

Forward School

Why is Big Data important?

Forward School

Domain Specific Examples of Big Data

- Typical domains where BDA can be applied to
  - Businesses such as industry, retail, logistics, agriculture
  - But also - cities, homes, environment, energy systems, healthcare
  - Internet of things too!

Forward School

Domain specific example: Web

- Web analytics
- Performance Monitoring
- Ad Targeting and Analytics
- Content recommendation (delivery or creation)

Forward School

Domain specific example: Financial

- Credit Risk Modeling
- Fraud Detection

Forward School

Domain specific example: Healthcare

- Patient similarity-based decision intelligence Application
- Epidemiological Surveillance
- Adverse Drug Events Prediction
- Detecting Claim Anomalies
- Evidence-based Medicine
- Real-time health monitoring

Forward School

Domain specific example: IoT

- Intrusion Detection
- Smart Parking
- Smart Roads
- Structural Health Monitoring
- Smart Irrigation

Forward School

Domain specific example: Environment

- Weather monitoring
- Air Pollution Monitoring
- Noise Pollution Monitoring
- Forest fire detection
- River flooding detection
- Water quality monitoring

Forward School

Domain specific example: Logistics and Transportation

- Real-time fleet tracking
- Shipment Monitoring
- Remote Vehicle Diagnostics
- Route generation and scheduling
- Hyper-local delivery
- Cab/taxi aggregators

Forward School

Domain specific example: Industry

- Machine diagnosis and prognosis
- Risk analysis of industrial operations
- Production planning and control

Forward School

Domain specific example: Retail

- Inventory management
- Customer Recommendations
- Store layout optimization
- Forecasting demand

Forward School

Analytics Flow for BDA

Forward School

Analytics Flow for BDA

- Data collection
- Data Preparation
- Analysis Type
- Analysis Modes
- Visualization

Forward School

Analytics Flow for BDA

- Data collection
  - The first step for any analytics application
  - Collection and ingestion into a big data stack first before any analysis can be started
  - The choice of tools and framework for collection
    - Depends on the source of data and the type of data being ingested
    - Usable connectors for data collections such as publish-subscribe messaging framework, messaging queues, source-sink connectors, database connectors etc.

- Ingestion: https://www.quora.com/What-is-the-difference-between-data-ingestion-and-data-integration-Are-these-terms-general-synonymous-or-do-they-refer-to-different-concepts

Forward School

Data preparation

- Data might not be directly usable
  - Dirty, noisy, corrupt, missing values, duplicates, inconsistent abbreviations, inconsistent units, typos, spellings/formatting etc.
  - Need to be resolved before it can be processed
- Data preparation includes
  - Cleansing, wrangling/mungling, de-duplication, normalization, sampling and filtering

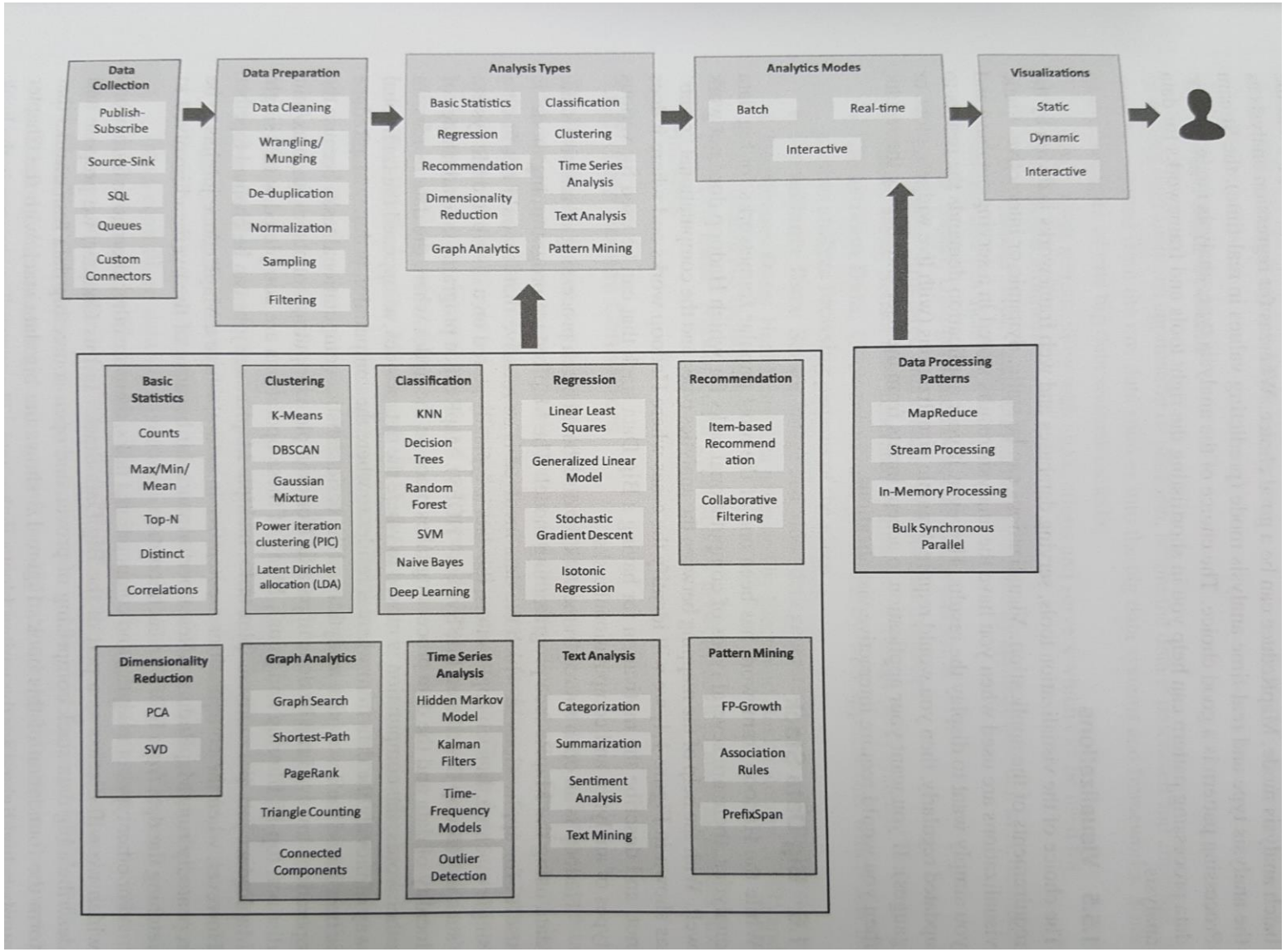
Forward School

Data preparation

- Data Cleansing
  - Detects and resolves issues such as corrupt records, records with missing values, records with bad formatting.
- Data wrangling/munging
  - Transforming data from one raw format to another
  - E.g. inconsistencies between multiple text files with different field separators.
    - Data wrangling resolve this by parsing the raw data from different sources and transforming it into one consistent format
- Normalization
  - Is required when data from different sources use different units or scale or have different abbreviations for the same thing
  - E.g. Weather data report by some stations in Celsius while the other stations may use Fahrenheit.
- Filtering and sampling
  - Useful when we want to process only data that meets certain rules.
  - Filtering can also be useful to reject bad records with incorrect or out-of-range values

Forward School

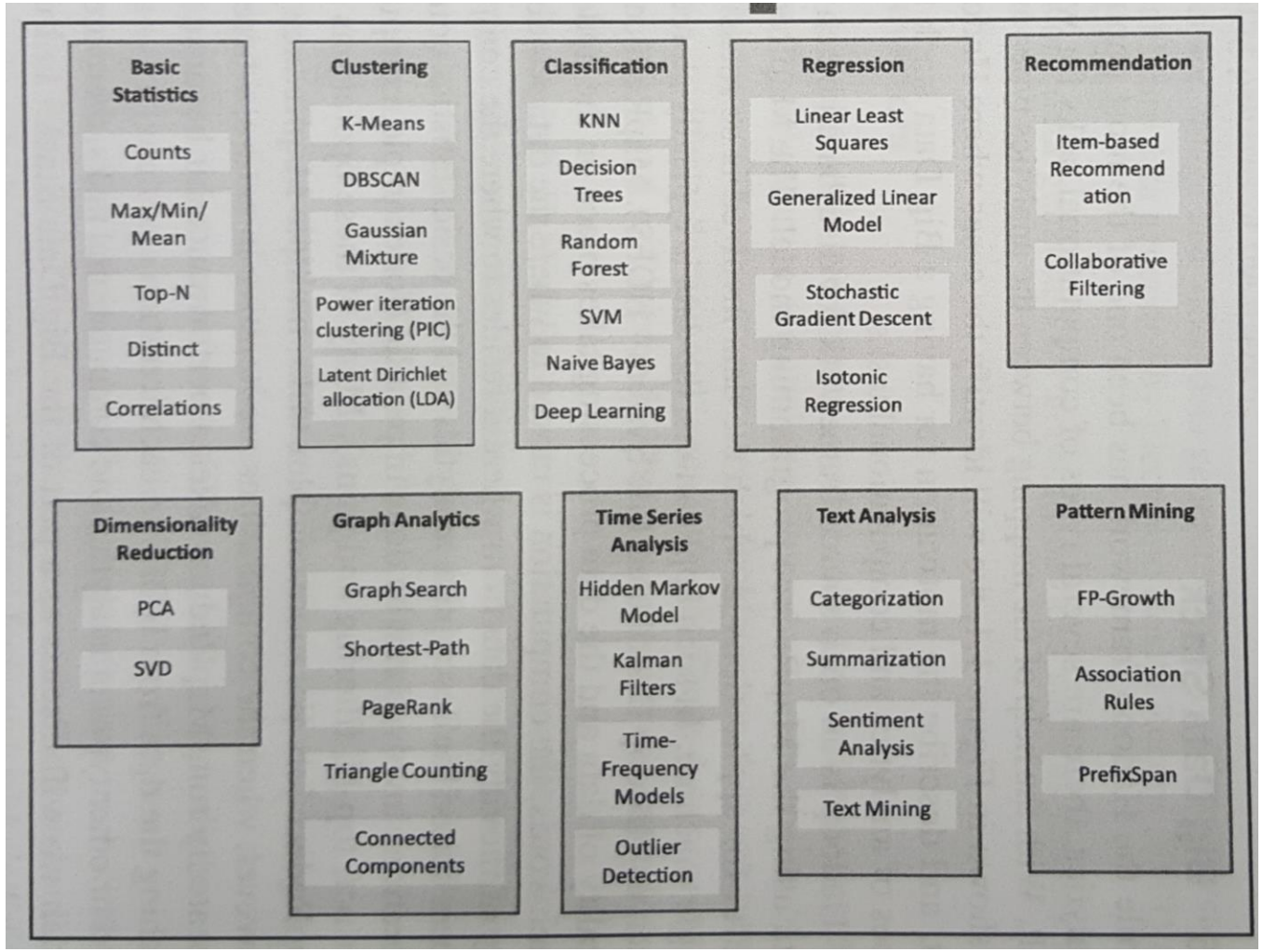
Analysis Types and modes



Forward School

Analysis Types

- Once data is collected and prepared, one can select appropriate types of analysis to analyse the data
- Normally Big data tools and framework will be selected to execute appropriate algorithms



Forward School

Analysis modes

- Three modes of analysis
  - Batch, real-time or interactive
- Choice – depends on the requirements of the application
  - Real-time mode - If results are required to be updated after short interval of time (say every few seconds
  - Batch mode – if results are required to be generated and updated on larger timescales (daily or monthly)
  - Interactive mode – if results should be generated via flexible query of data on demand

Forward School

Analysis mode

- Once the choice of analysis type and mode is made, one can determine the data processing pattern that can be used
- E.g. –
  - MapReduce can be a good choice if basic statistics (type) and batch analysis (mode) are selected
  - Stream processing pattern is a good choice if Regression (type) and real-time analysis (mode) are required
- This is the shortlisting/selection process based on types and modes to come to the suitable choice of tools and frameworks for data analysis

Forward School

Visualization

- The choice of the visualization tools, serving databases and web framework is driven by the requirements of the applications
- Visualization can be
  - Static – to only display the results of an analysis
  - Dynamic – to display regular update of results through live widget, plot etc.
  - Interactive – to change display of results based on inputs from the user

Forward School

Big Data Stack

Forward School

Big Data Stack

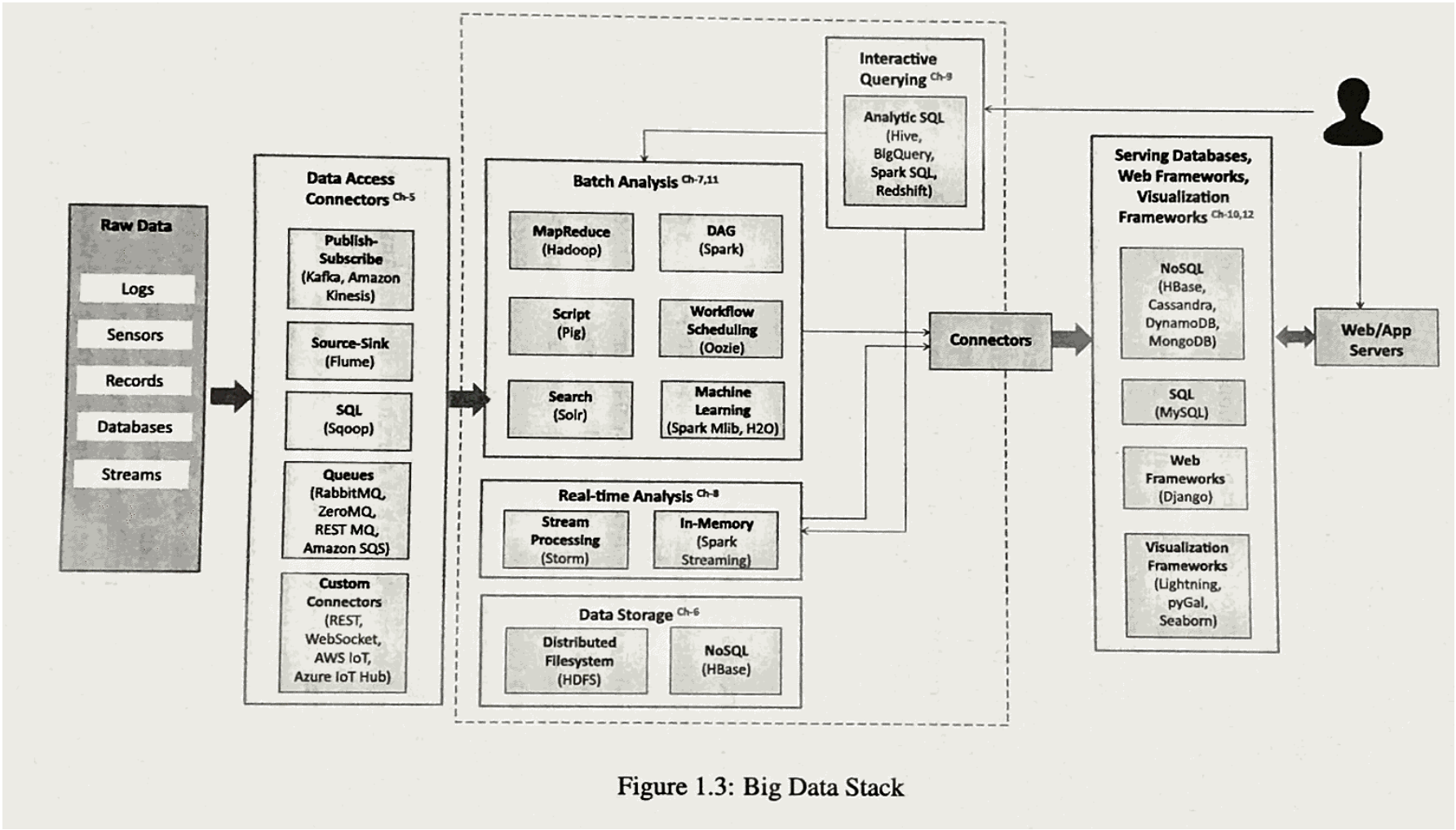


Figure 1.3: Big Data Stack

Forward School

Big data stack

- Hadoop framework is one of the most popular frameworks for big data analytics
- Can you name other tools/frameworks?
  - https://www.kdnuggets.com/2016/03/top-big-data-processing-frameworks.html

Forward School

Raw Data Sources

- This is where data collection takes place
- Raw data sources include
  - Logs – logs generated by web applications and services which can be used for performance monitoring
  - Transactional data – Transactional data generated by applications such as eCommerce, Banking and Financial
  - Social Media – Data generated by social media platform
  - Databases – Structured data residing in relational databases
  - Sensor Data – sensor data generated by Internet of Things (IoT) systems
  - Clickstream data – clickstream data generated by web applications which can be used to analyse browsing patterns of the users
  - Surveillance Data: Sensor, image and video data generated by surveillance systems
  - Healthcare data – healthcare data generated by Electronic Health Record (EHR) and other healthcare applications
  - Network Data – Network data generated by network devices such as routers and firewalls

Data Access Connectors

- Include tools and frameworks for collecting and ingesting data from various sources into the big data storage and analytics frameworks
- Choice of connects -> driven by the type of the data source
- Example connectors
  - Publish-subscribe messaging
    - a communication model that involves publishers (source of data), brokers (manager of the data subscription and sending) and consumers (recipient of the data).
    - Example – Apache Kafka, Amazon Kinesis
  - Source-sink connectors
    - It allows efficiently collecting, aggregating and moving data from various sources (server logs, databases, social media, streaming sensor data etc.) into a centralized data source (such as a distributed file system).
    - Example – Apache Flume

Forward School

Data Access Connectors (cont.)

- Examples
  - Database connectors
    - Mainly tools that allow import of data from relational database management systems into big data storage and analytics frameworks for analysis
    - Example – Apache Sqoop
  - Messaging queues
    - They are useful for push-pull messaging, where the producers push data to the queues and the consumers pull the data from the queues
    - Producers and consumers do not need to be aware of each other
    - Example – RabbitMQ, ZeroMQ, RestMQ and Amazon SQS
  - Custom Connectors
    - They are built based on the source of the data and the data collection requirements
    - Some examples include – custom connectors for collecting data from social network, NoSQL databases and IoT
    - Example technology/tools – custom connector based on REST, WebSocket, MQTT, IOT connectors such as Amazon IoT, Azure IoT Hub

Data Storage

- The data storage block in the big data stack includes distributed filesystems and non-relational (NoSQL) databases
  - Store data collected from the raw data sources using data access connectors
- Examples
  - Hadoop Distributed File System (HDFS)
    - a distributed file system that runs on large cluster and provides high-throughput access to data
    - Data stored in HDFS can be analyzed with various big data analytics frameworks built on top of HDFS
  - HBase
    - Is a scalable, non-relational, distributed, column-oriented database
    - A NoSQL database that provides structured data storage for large tables

Forward School

Batch Analytics

- Tools and frameworks include
  - Hadoop-MapReduce – MapReduce is the programming model used to develop batch analysis jobs to be executed in Hadoop clusters
  - Pig – Pig is a high-level data processing language which makes it easy for developers to write data analysis scripts which are translated into MapReduce programs by the Pig compiler
  - Oozie – a workflow scheduler system that allows managing Hadoop jobs. One can create workflows which are a collection of actions (such as MapReduce jobs) arranged as Direct Acyclic Graphs (DAG)
  - Spark – Apache Spark is an open source cluster computing framework for data analytics. It includes high-level tools for data analysis such as Spark Streaming, Spark SQL, Mllib and GraphX.
  - Solr – Apache Solr is a scalable and open-source framework for searching data.
  - Machine Learning – There are many ML libraries or framework that work on top of a big data framework, such as Mllib or H20.

Real-time analytics

- The real-time analytics block includes the Apache Storm and Spark Streaming frameworks
- Apache Storm
  - A framework for distributed and fault-tolerant real-time computation
  - Can be used for real-time processing of streams of data
  - Can consume data from a variety of sources such as publish-subscribe messaging frameworks (Kafka, Kinesis), messaging queues (RabbitMQ, ZeroMQ) and other custom connectors
- Spark Streaming
  - A component of Spark which allows analysis of streaming data such as sensor data, clickstream data, web server logs
  - The streaming data is ingested and analyzed in micro-batches
  - Spark Streaming enables scalable, high throughput and fault-tolerant stream processing

Forward School

Interactive Querying

- Interactive querying systems allow users to query data by writing statements in SQL-like languages
- Examples
  - Spark SQL – a component of Spark which enables interactive querying. Useful for querying structured and semi-structured data using SQL-like queries
  - Hive – Apache Hive is a data warehousing framework built on top of Hadoop. Hive Query Language is an SQL-like query language for querying data residing in HDFS
  - Amazon Redshift – is a fast, massive-scale managed data warehouse service. It specializes in handling queries on datasets of sizes up to a petabyte or more parallelizing the SQL queries across all resources in the Redshift cluster
  - Google BigQuery – is a service for querying massive datasets, allows querying datasets using SQL-like queries

Serving Databases, Web and Visualization Frameworks

- The results from the various analytics blocks (processing, analysis etc.) are stored in serving databases for further/subsequent tasks
  - Presentation and visualization
- These serving databases will allow the analyzed data to be queried and presented in (web) applications
- Some examples
  - MySQL – one of the most widely used Relational Database Management System (RDBMS) and is a good choice to store structured data as a serving database for data analytics
  - Amazon DynamoDB – a fully managed, scalable, high-performance NoSQL database service from Amazon. It can be an excellent choice when it comes to storing and retrieving any amount of data and the ability to scale up or down the provisioned throughput
  - Cassandra – is a scalable, highly available, fault tolerant open souce non-relational database system
  - MongoDB – a document oriented non-relational database system. It is powerful, flexible and highly scalable databased designed for web applications. It can be a good choice as a serving database for data analytics application.

Serving Databases, Web and Visualization Frameworks

- Web frameworks – example Django
  - Python based open-source web application framework.
  - Model-Template-View architecture
  - Separation of the data model from business rules and user interface
- Visualization framework/tools
  - Lightning – a framework for creating web-based interactive visualizations
  - Pygal – an easy python-based library for charting purposes
  - Seaborn – another python-based visualization library for plotting attractive statistical plots

Forward School

Big Data Stack

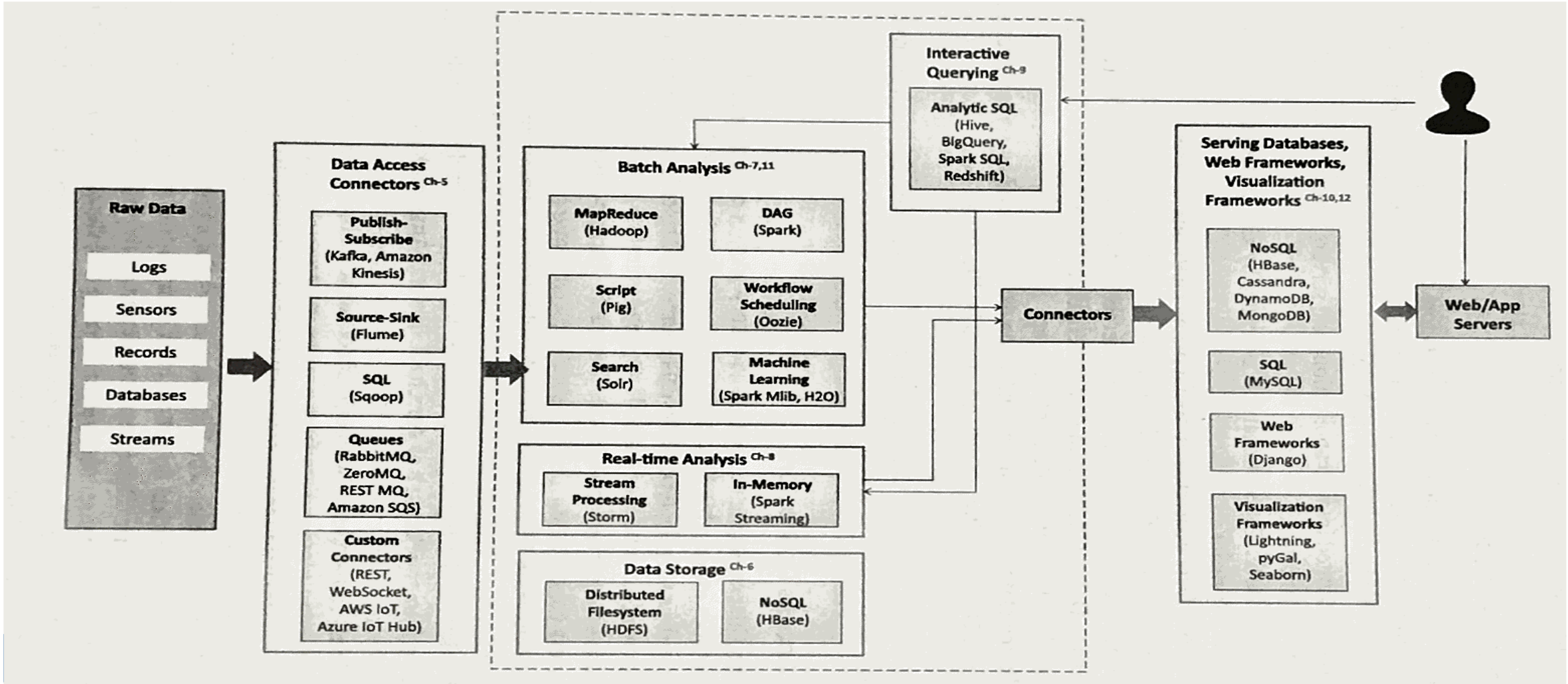


Figure 1.3: Big Data Stack