Big Data Analytics for Medical Imaging

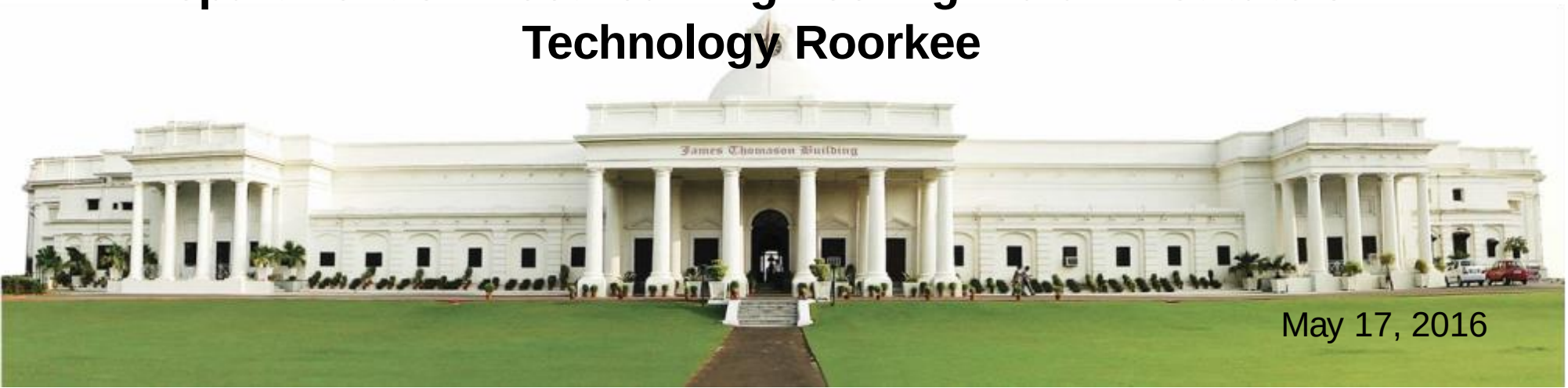
a seminar presentation by

UTPAL KANT

Under the guidance of

DR. VINOD KUMAR
&
DR. P. SUMATHI

Department of Electrical Engineering Indian Institute of Technology Roorkee



May 17, 2016

Table of Contents

Introduction to Big Data

- A collection of large and complex data sets which are difficult to process using common database management tools or traditional data processing applications.

- 5 V'S Characteristics of Big Data
  - Volume
  - Velocity
  - Variety
  - Veracity
  - Value

Hadoop Framework

Apache Hadoop is an open source software framework for storage and large scale processing of data-sets on clusters of commodity hardware.

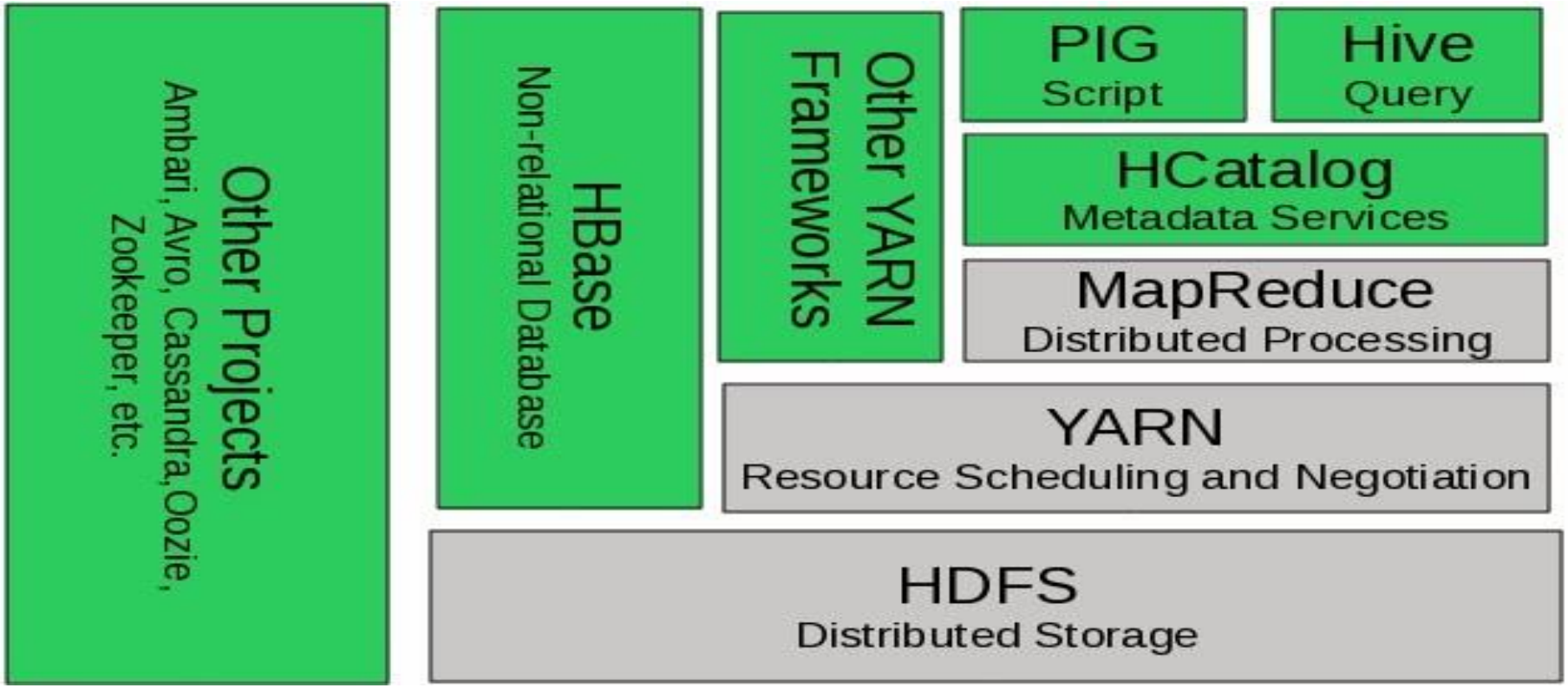


Image Source: [7]

Hadoop Distributed File System

HDFS

Distributed, scalable, and portable file- system written in Java for the Hadoop framework

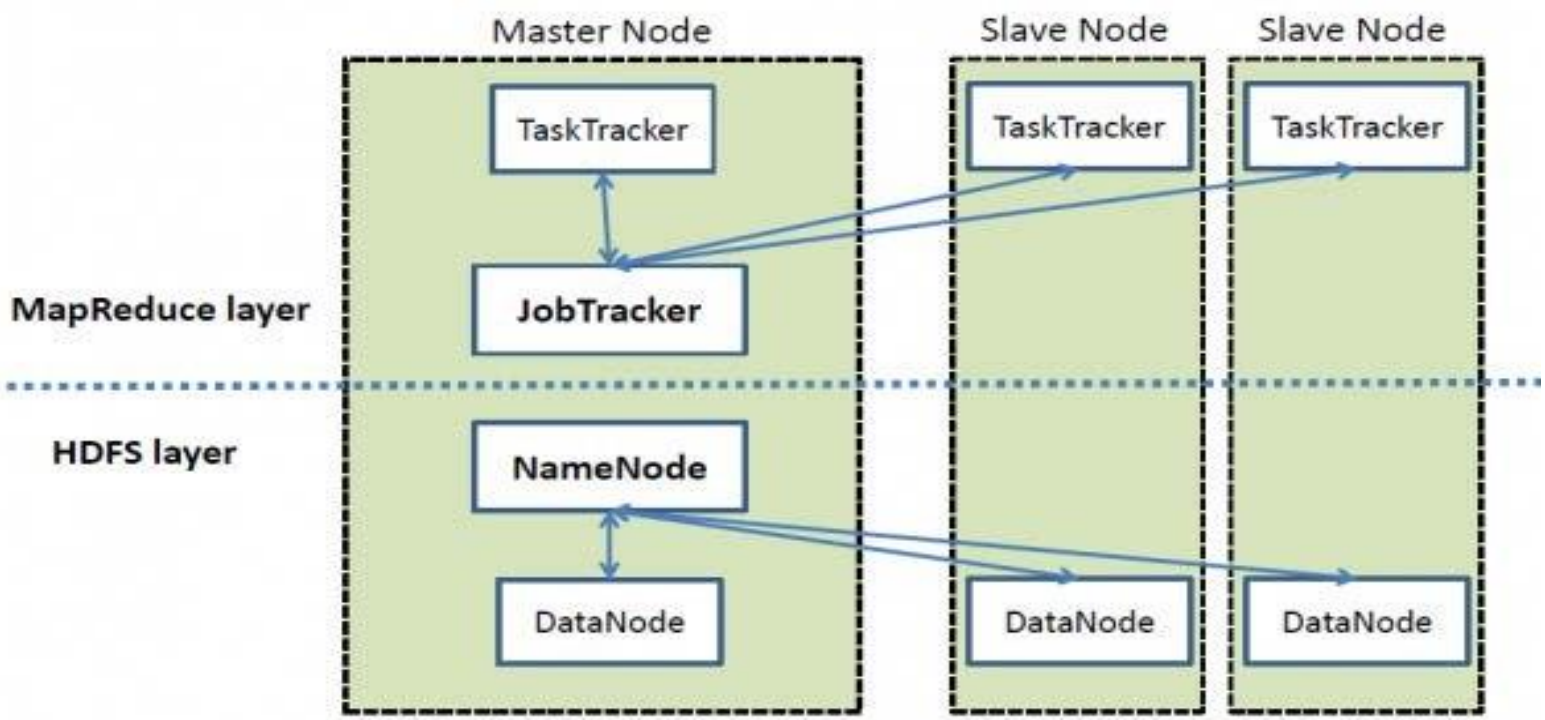


Image Source: [8]

HDFS

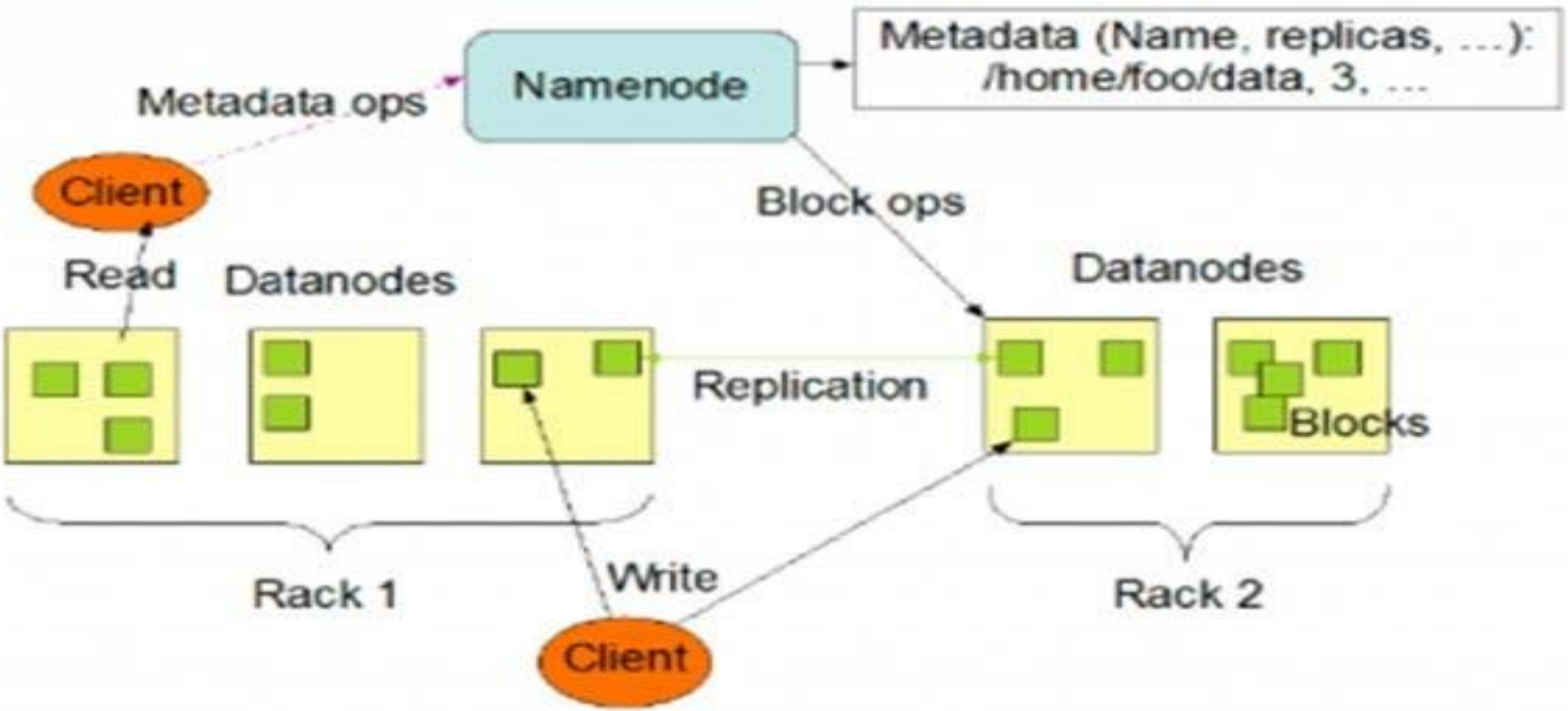


Image Source: [8]

HDFS

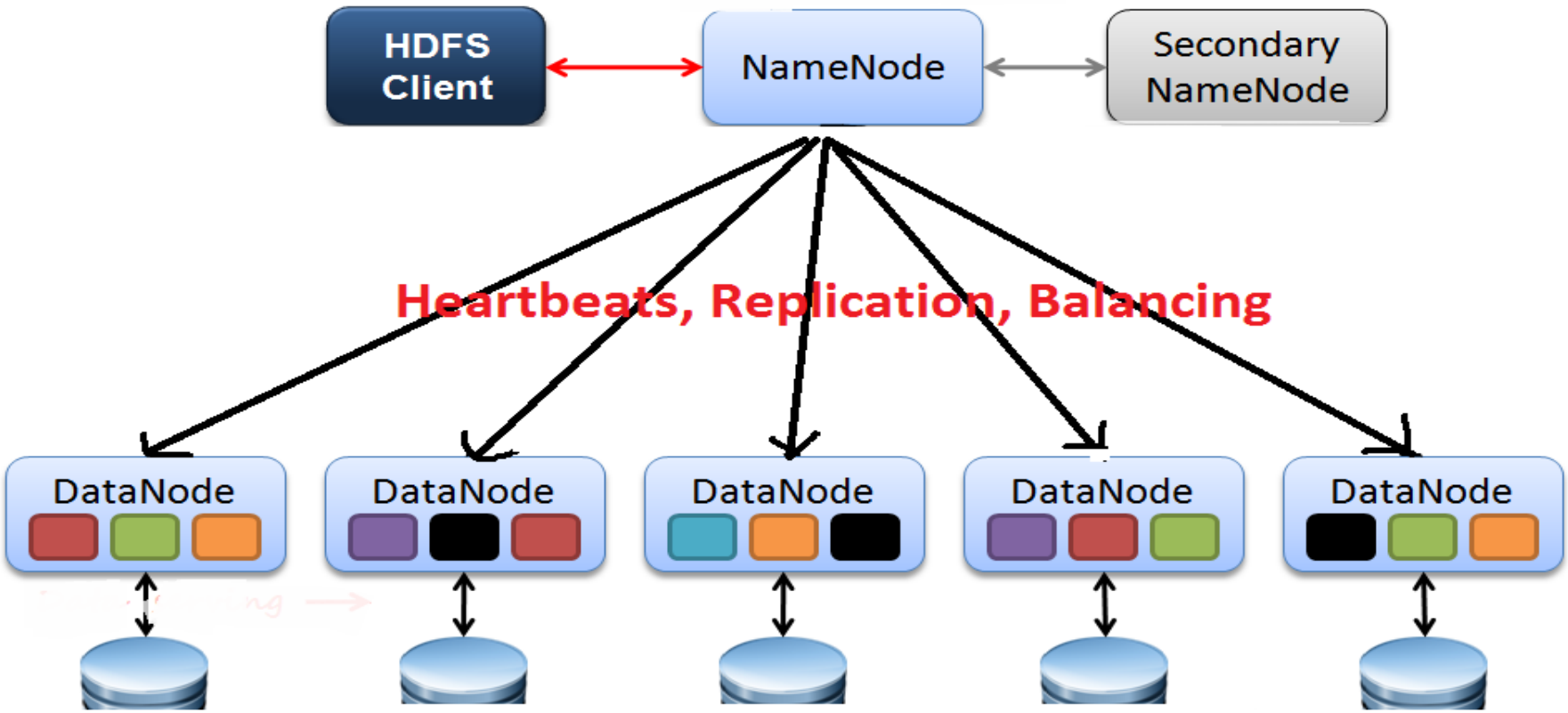


Image Source: 8

MapReduce Engine

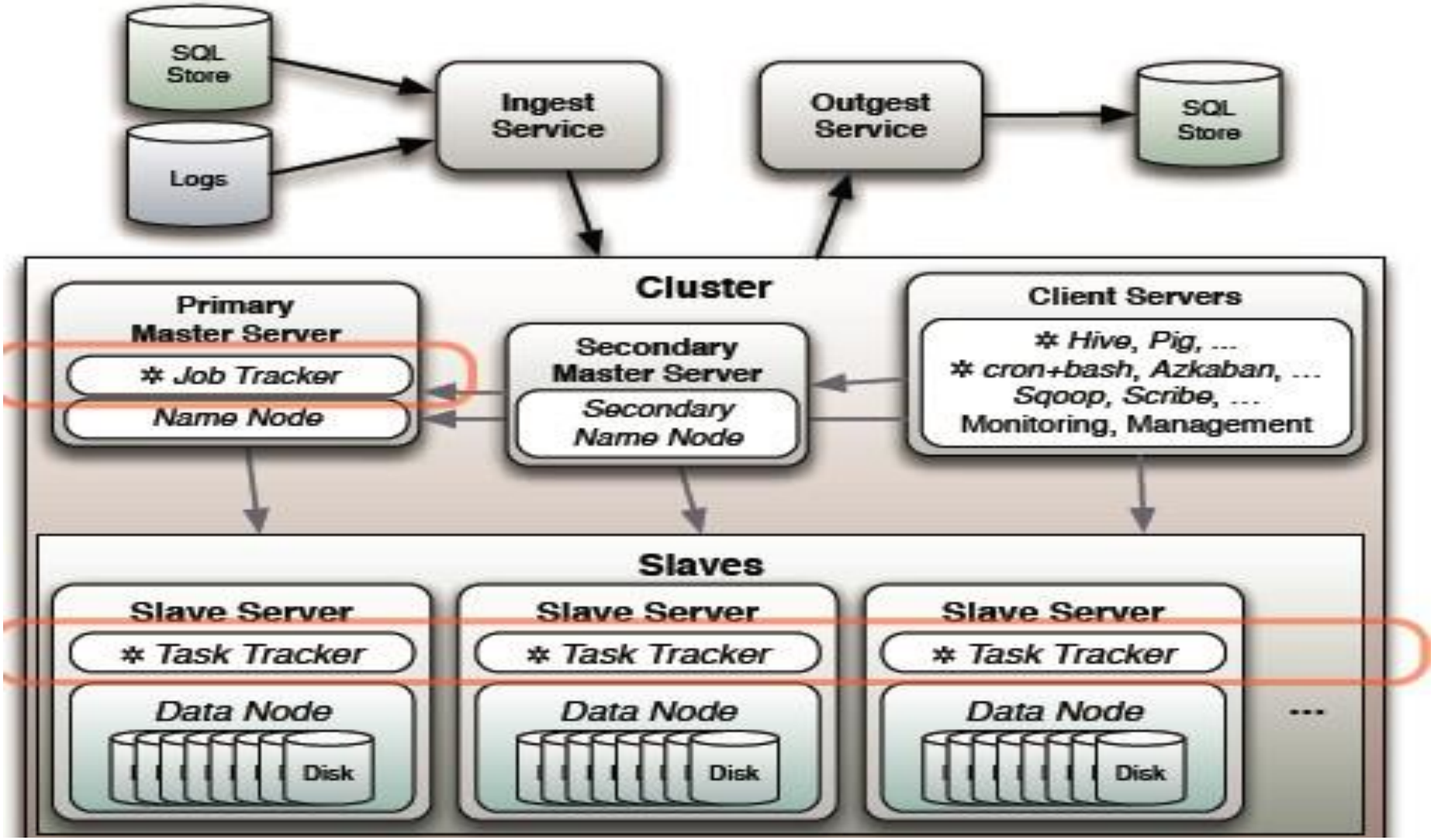
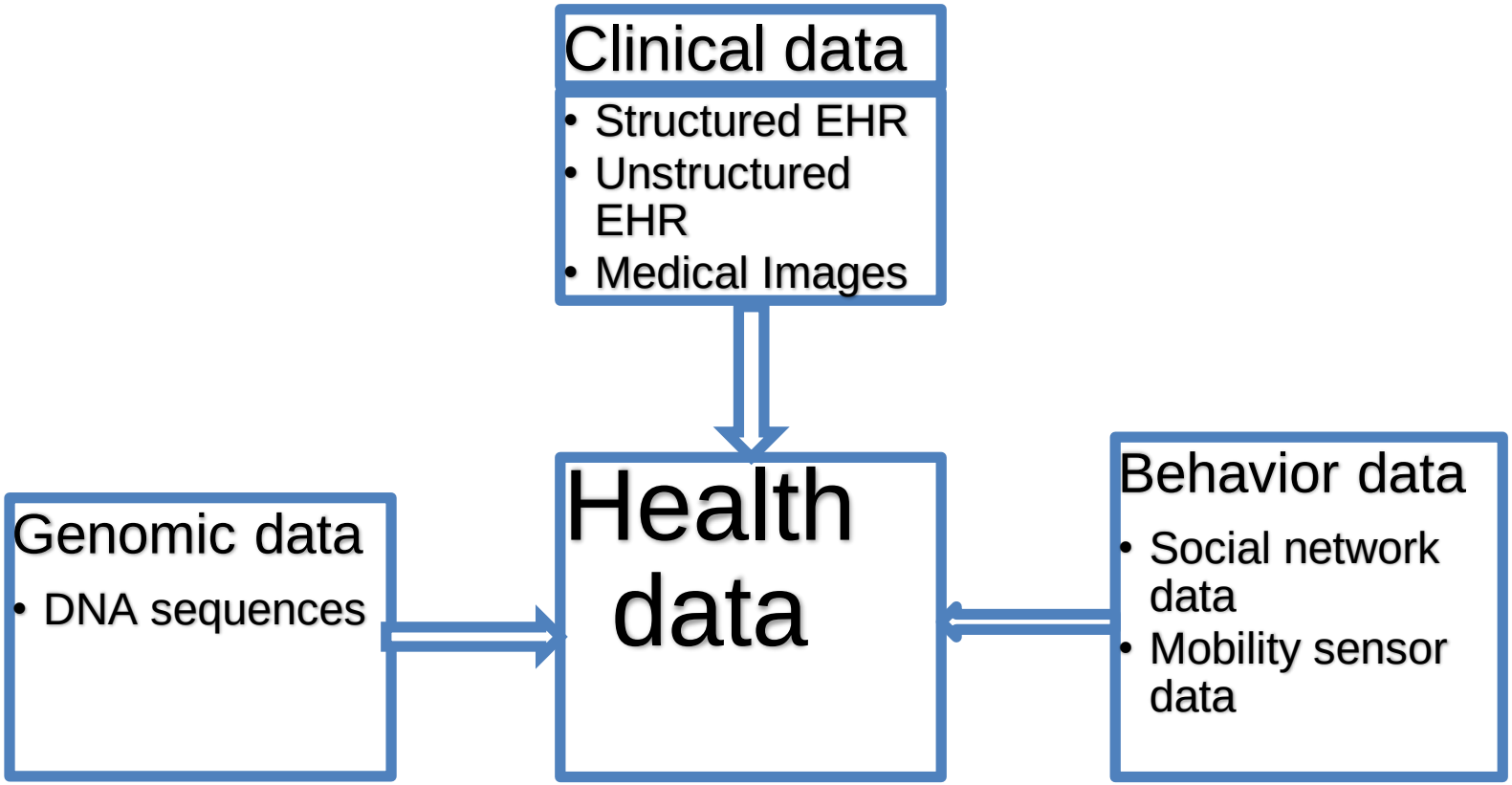
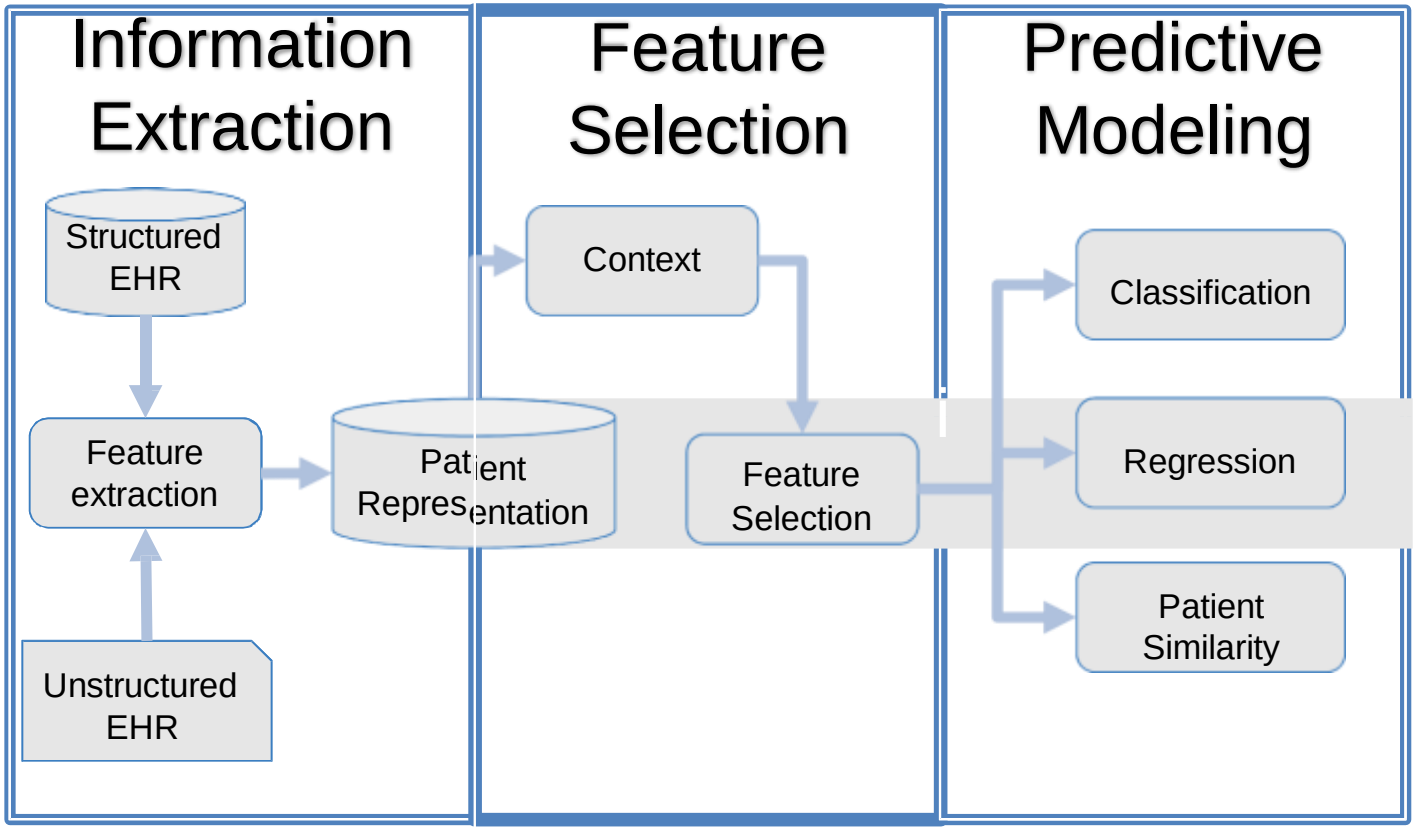


Image Source: [8]

Big Data in Health Care



Analytic Platform



Text Mining in Healthcare

- Text mining
  - Information Extraction
    - Name Entity Recognition
  - Information Retrieval
- Clinical text vs. Biomedical text
  - Biomedical text: medical literatures (well-written medical text)
  - Clinical text is written by clinicians in the clinical settings

Combining Knowledge- and Data-driven Risk Factors

Prediction Models

- ❖ Continuous outcome: Regression and Classification
- ❖ Survival outcome: Hazard Regression
- ❖ Patient Similarity

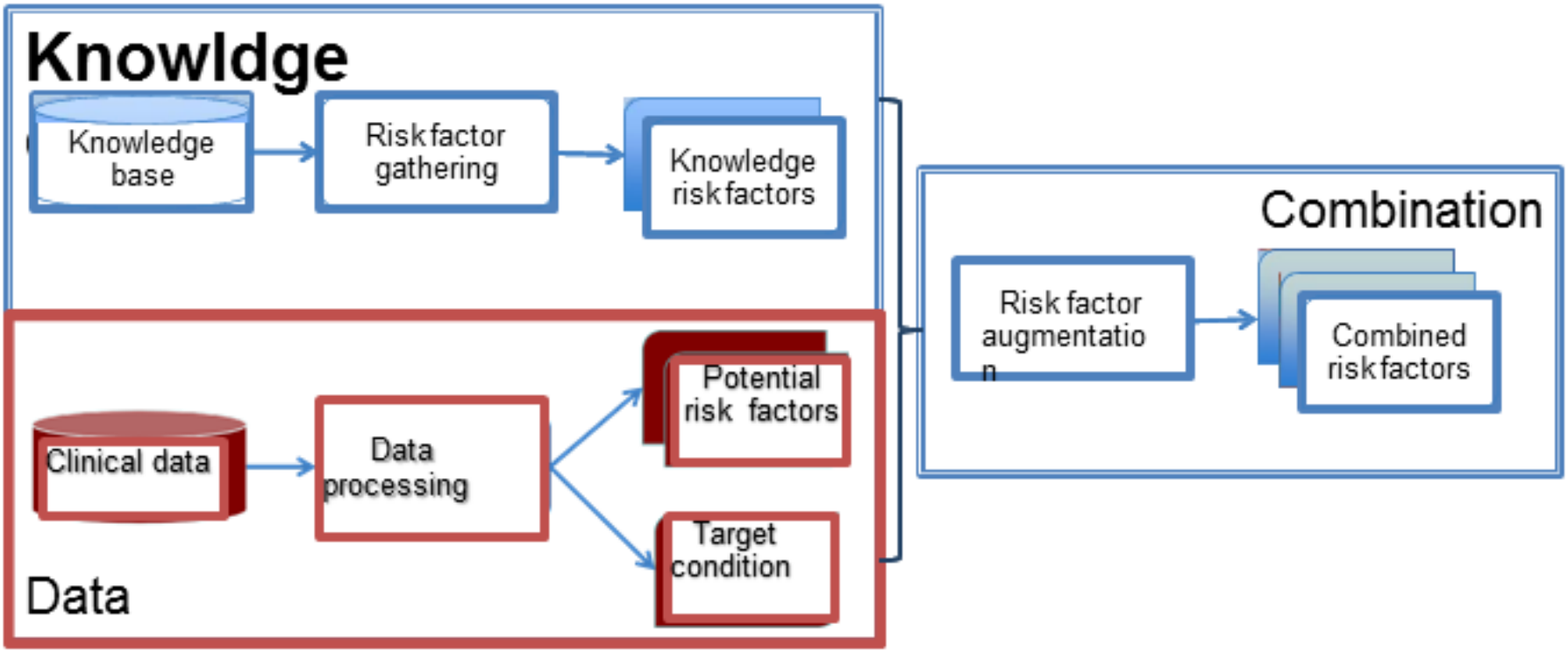


Image Data is Big !!!

- By 2015, the average hospital will have two-thirds of a *petabyte* (665 terabytes) of patient data, 80% of which will be unstructured image data like CT scans and X-rays.
- Medical Imaging archives are increasing by 20%-40%
- PACS (Picture Archival & Communication Systems) system is used for storage and retrieval of the images.

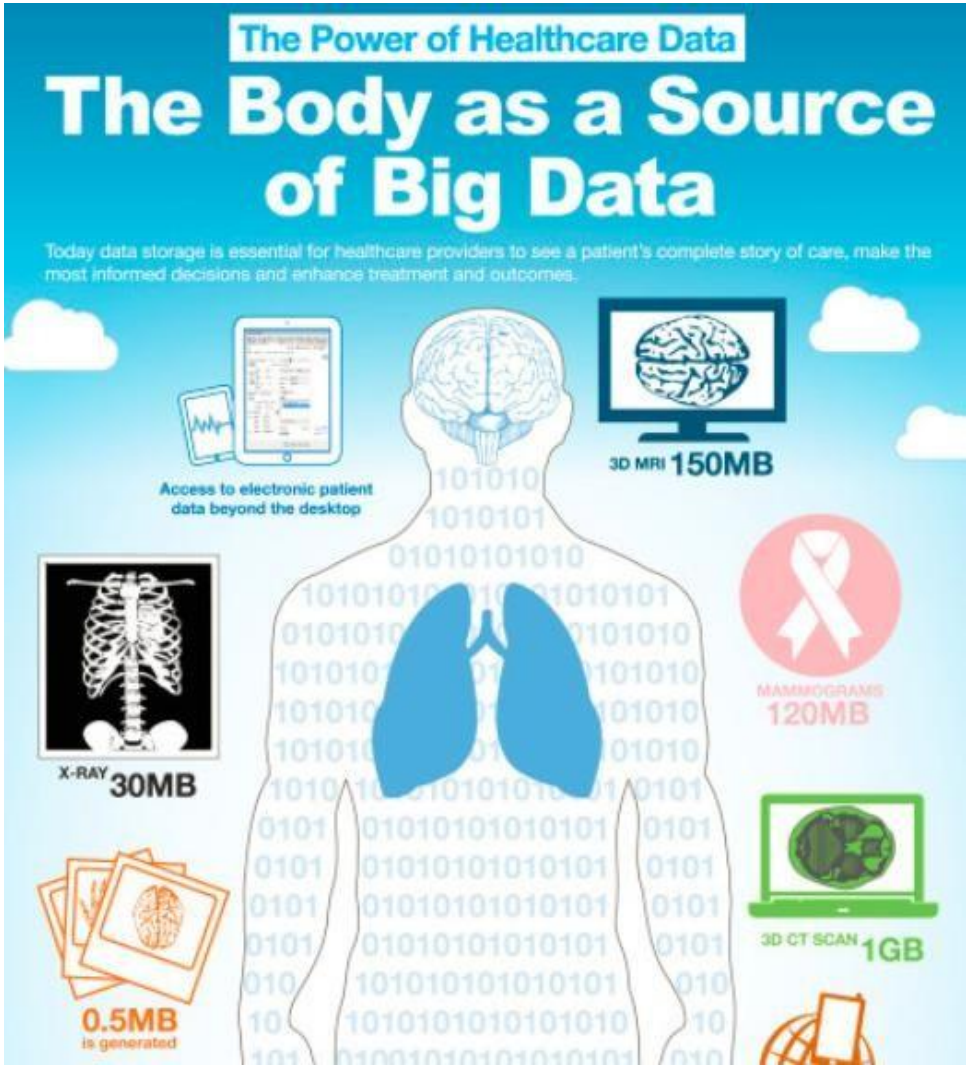
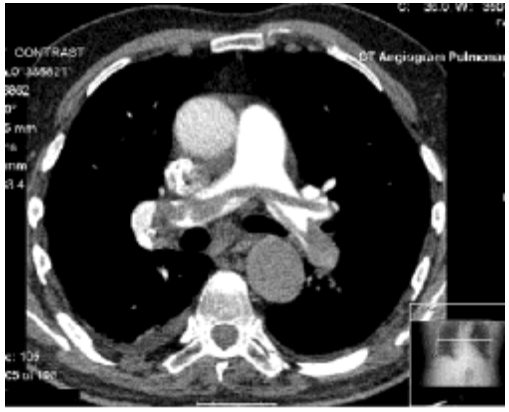
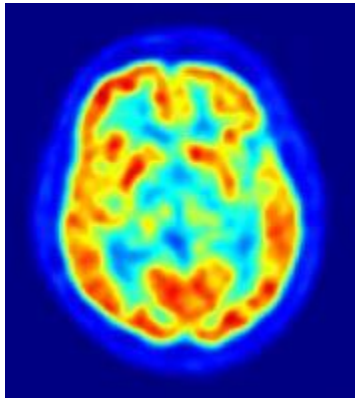


Image Source: [8]

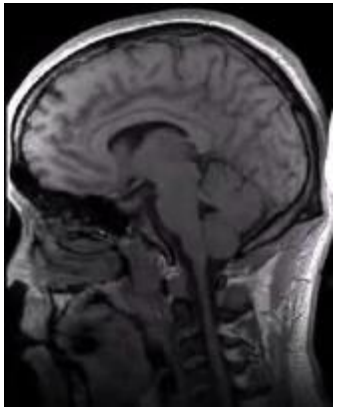
Popular Imaging Modalities in Healthcare Domain



Computed Tomography (CT)



Positron Emission Tomography (PET)



Magnetic Resonance Imaging (MRI)

- The main challenge with the image data is that it is not only huge, but is also high-dimensional and complex.
- Extraction of the important and relevant features is a daunting task.
- Many research works applied image features to extract the most relevant images for a given query.

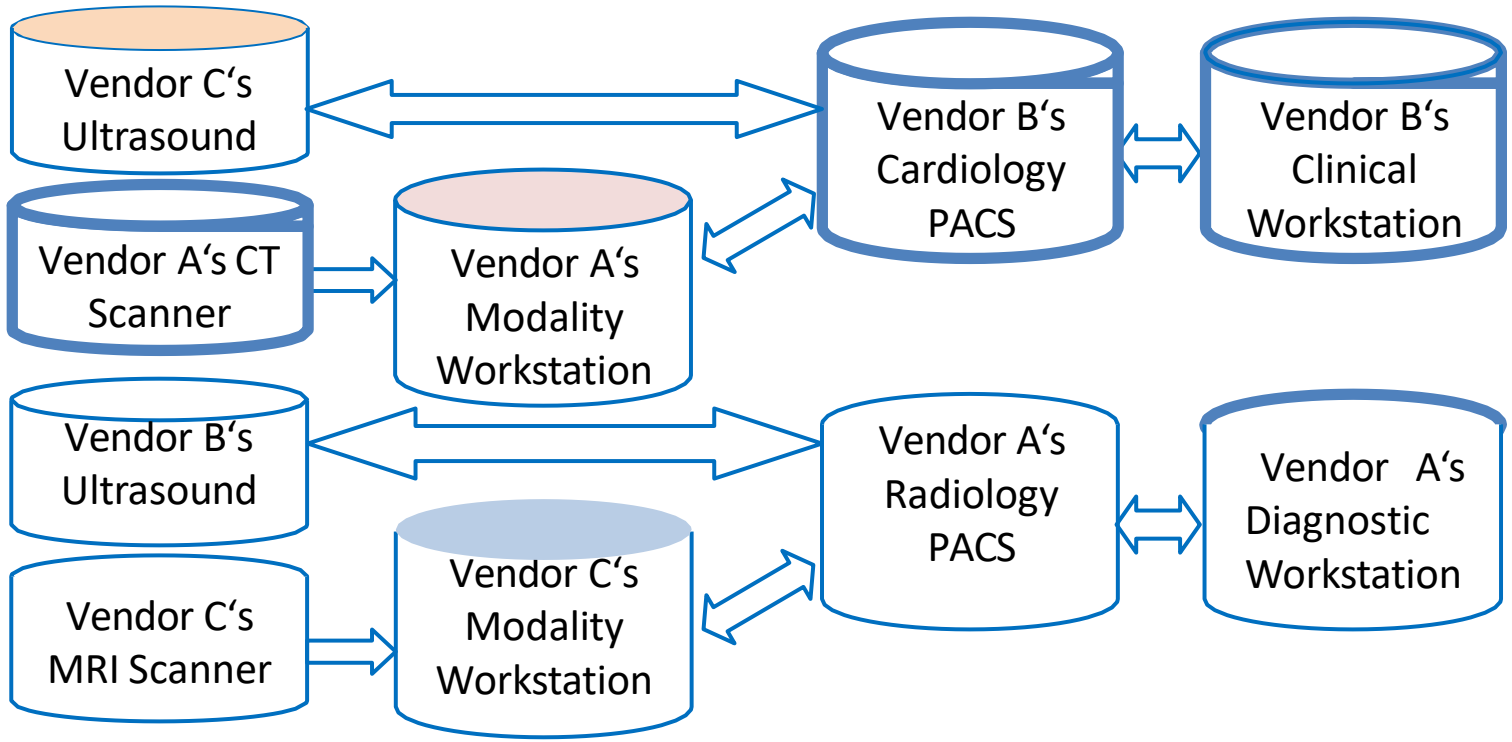
Image Source: wikipedia

Evolution of Medical Imaging Informatics

- Radiology invented the concept of PACS (Picture Archival & Communication Systems)
  - "Father of PACS" - The late Samuel J. Dwyer, III, PhD
- It is a solution that is born out of real-world needs
  - Due to a **need** to **improve** diagnostic capabilities
  - These needs are so effectively fulfilled that PACS these days are no longer limited to only medical images nor strictly for the radiology discipline
- PACS (next to the EMR) is to be one of the most significant
  - clinical information systems in the healthcare enterprise

Typical Multi-PACS Environment

- Typical departmental PACS implementation for Radiology (Post Digital Imaging and Communications in Medicine DICOM 3.0)
- Hospital with Radiology & Cardiology PACS



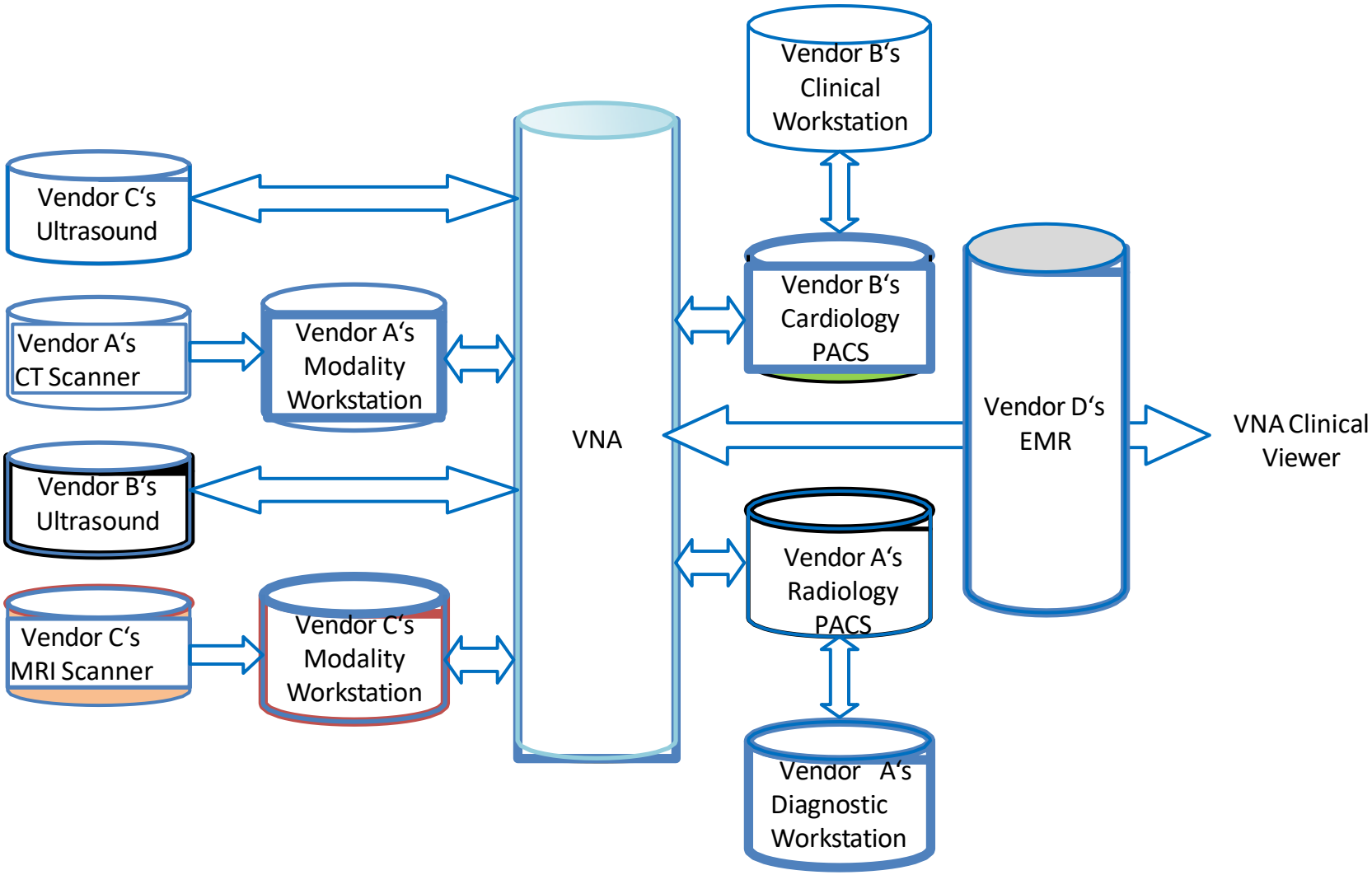
Problems with "Traditional" Implementation

- The "Traditional" implementation resulted in a "Silo model"
  - Such PACS silos were appropriate for departmental implementations
- As health IT adoption moves onto the enterprise levels, "Silo model" no longer serves clinical or operational needs
  - From a Technologist, Specialist, Clinician, Administrative, Technical and most importantly Patient's perspective
  - In the modern healthcare enterprise, we need a patient centric workflow

VNA Architecture

- Vendor Neutral Archive (VNA) Architecture
- Helps rectify inherent issues with DICOM
  - Private Tags
  - Transfer Syntax
  - Data Migration
- Enables single Web Client access across all

VNA Architecture



Cloudify the Imaging Grid

- The VNA Architecture has enabled solution architects "pushed" the limits a little further. towards the clouds

- Extending both the benefits of Cloud and VNA Architecture
  - This actually makes perfect sense
    - Vendor Neutrality for Interoperability
    - Facilitating Patient Centric Care
    - Cost by a need to use

- Gives new meaning to the phase
  - “Imaging Anytime, Anywhere, Anyplace”

Publicly Available Medical Image Repositories

Image database Name	Modalities	No. Of patients	No. Of Images	Size Of Data	Notes/Applications	Download Link
Cancer Imaging Archive Database	CT DX CR	1010	244,527	241GB	Lesion Detection and classification, Accelerated Diagnostic Image Decision, Quantitative image assessment of drug response	https://public.cancerimagingarchive.net/ncia/dataBasketDisplay.jsf
Digital Mammography database	DX	2620	9,428	211GB	Research in Development of Computer Algorithm to aid in screening	http://marathon.csee.usf.edu/Mammography/Database.html
Image CLEF Database	PET CT MRI US	unknown	306,549	316GB	Modality Classification , Visual Image Annotation , Scientific Multimedia Data Management	http://www.imageclef.org/2013/medical

Image CLEF Data

- ImageCLEF aims to provide an evaluation forum for the cross– language annotation and retrieval of images (launched in 2003)
- Statistics of this database :

  With more than 300,000 (in .JPEG format), the total size of the database > 300 GB

  contains PET, CT, MRI, and Ultrasound images

- Three Tasks

  Modality classification

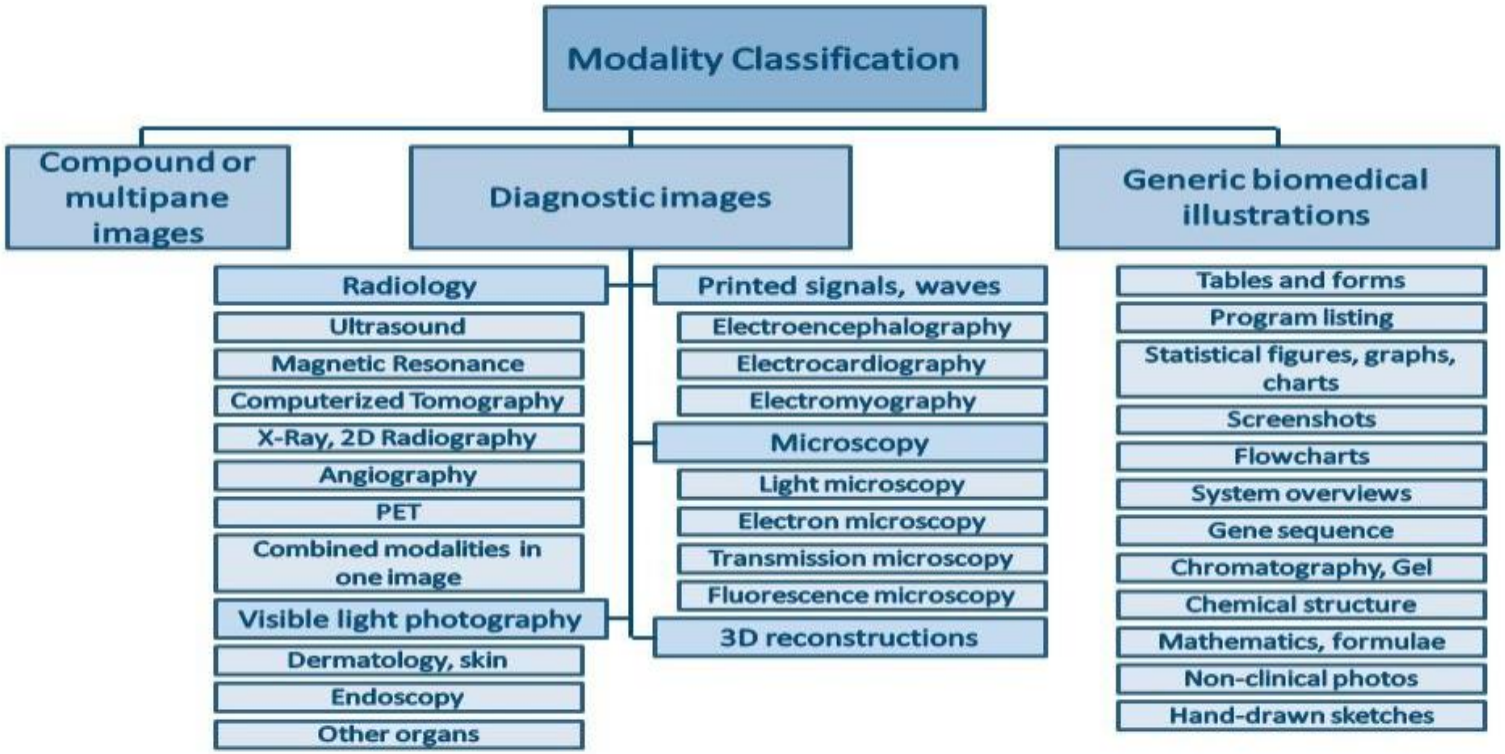
  Image–based retrieval

  Case–based retrieval

Medical Image Database available at http://www.imageclef.org/2013/medical

Why BDA in Medical Imaging?

Modality Classification Task



Modality is one of the most important filters that clinicians would like to be able to limit their search.

Medical Image Features

Photo-metric features exploit color and texture cues and they are derived directly from raw pixel intensities.

Geometric features: cues such as edges, contours, joints, polylines, and polygonal regions.

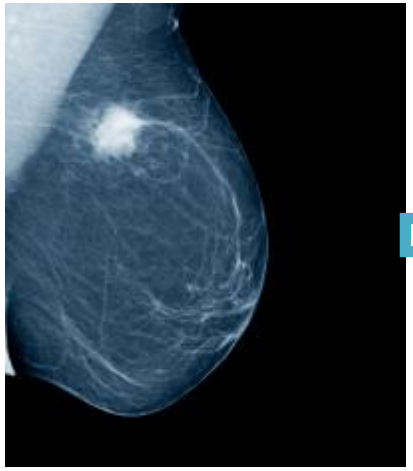
A suitable shape representation should be extracted from the pixel intensity information by region-of interest detection, segmentation, and grouping. Due to these difficulties, geometric features are not widely used.

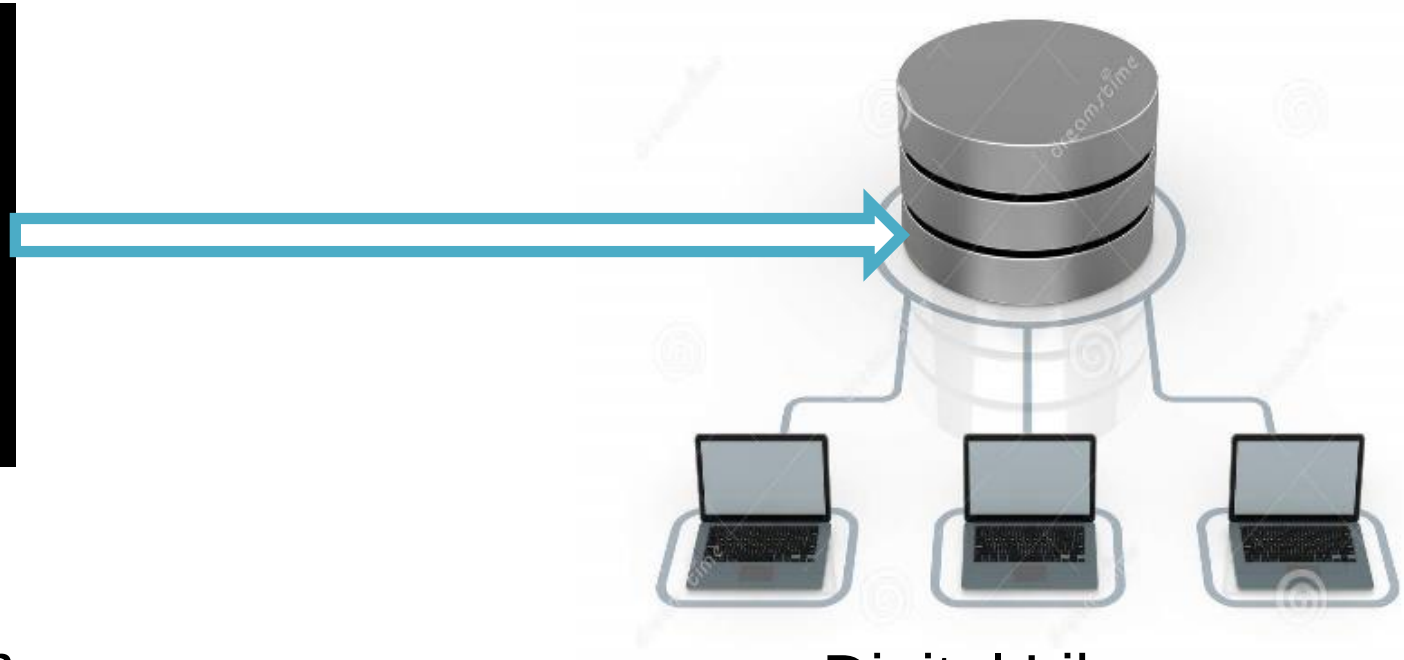
A Summary of Image Features/Descriptors Used in the Medical Domain

Category	Representations/cues	Examples
Photometric	Grayscale and color	Histograms[13,16]
		Moments[21,24]
		Block-based[17,19]
	Texture	Texture co-occurrence[16,20,21,23,24]
		Fourier power spectrum[21]
		Gabor features[15,20]
		Wavelet-based[14]
		Haralick's statistical features[32]
		Tamura features[18]
		Multiresolution autoregressive model[13]
Geometric	Point sets	Shape spaces[33]
	Contours/curves	Polygon approximation[34]
		Edge histograms[16,24,32]
		Fourier-based[13,16,34]
		Curvature scale space[35]
	Surfaces	Level sets/distance transforms[20,36]
		Gaussian random fields[37]
	Regions and parts	Statistical anatomical parts model[38]
		Wavelet-based region descriptors[39]
		Spatial distributions of ROIs[40]
	Other	Global shape (size, eccentricity, etc.)[16,17]
		Morphological[20,42,43]
		Location and spatial relationships[17,20]

Content-based Image Retrieval

Personalised Medicine: Big Data Supported Diagnosis





Mammogram

Digital Library

Content-based Image Retrieval

Two components

- Image features/descriptors - bridging the gap between the visual content and its numerical representation.
- These representations are designed to encode color and texture properties of the image, the spatial layout of objects, and various geometric shape characteristics of perceptually coherent structures.
- Assessment of similarities between image features based on mathematical analyses, which compare descriptors across different images.
- Vector affinity measures such as Euclidean distance, Mahalanobis distance, KL divergence, Earth Mover's distance are amongst the widely used ones.

Conclusion

- Big data analytics is a promising right direction which is in its infancy for the healthcare domain.
- Healthcare is a data-rich domain. As more and more data is being collected, there will be increasing demand for big data analytics.
- Unraveling the “Big Data” related complexities can provide many insights about making the right decisions at the right time for the ppersonalised medicine.
- Efficiently utilizing the colossal healthcare data repositories can yield some immediate returns in terms of patient outcomes and lowering care costs.

Thank You

References

1. Jeffrey Dean and Sanjay Ghemawat. “Mapreduce: simplified data processing on large clusters”, Communications *of the ACM,* vol. 51, no. 1, pp.107-113, 2008.
2. Marco Viceconti, Peter Hunter, and Rod Hose “Big Data, Big Knowledge: Big Data for Personalized Healthcare”, IEEE Journal of Biomedical And Health Informatics, vol. 19, no. 4, July 2015.
3. Alberto Bartesaghi, Guillermo Sapiro, and Sriram Subramaniam “An Energy-Based Three-Dimensional Segmentation Approach for the Quantitative Interpretation of Electron Tomograms”, IEEE Transactions on Image Processing, vol. 14, no. 9, Sept. 2005.
4. Chao-Tung Yang, Lung-Teng Chen, Wei-Li Chou, and Kuan-Chieh Wang. “Implementation of a medical image file accessing system on cloud computing”, *IEEE International Conference on Computational Science and Engineering (CSE),* pages 321-326. IEEE, 2010.
5. Carlos O R, Fernando L K, Carlos B W, Jorge W, Armando F, and Giovanni S S. “A cloud computing solution for patient's data collection in health care institutions”, *IEEE International Conference on eHealth, Telemedicine, and Social Medicine (ETELEMED),* pp 95-99, 2010.
6. Wenan Chen, Charles Cockrell, KR Ward, and Kayvan Najarian. “Intracranial pressure level prediction in traumatic brain injury by extracting features from multiple sources and using machine learning methods”, *IEEE International Conference on Bioinformatics and Biomedicine (BIBM),* pp. 510-515, 2010.
7. https://cloudera.com
8. https://hadoop.apache.

References

9. Antoine Widmer, Roger Schaer, Dimitrios Markonis, and Henning Müller. “Gesture interaction for content based medical image retrieval”, *In Proceedings of International Conference on Multimedia Retrieval,* pp. 503-506. ACM, 2014.
10. Konstantin Shvachko, Hairong Kuang, Sanjay Radia, and Robert Chansler. “The hadoop distributed file system” *IEEE Symposium on Mass Storage Systems and Technologies (MSST) ,* pp. 1-10. IEEE, 2010.
11. Dalia Sobhy, Yasser El-Sonbaty, and M Abou Elnasr. “Medcloud: healthcare cloud computing system”, *IEEE International Conference on Internet Technology And Secured Transactions,* pp. 161-166. IEEE, 2012.
12. Akgül, Ceyhun Burak, et al. "Content-based image retrieval in radiology: current status and future directions." *Journal of Digital Imaging,* vol. 24 No. 2 pp. 208-222, 2011.
13. Müller, Henning, et al. "A review of content-based image retrieval systems in medical applications-clinical benefits and future directions”, *International journal of medical informatics,* vol. 73 no.1 pp. 1-24, 2004.