A BIG DATA CURE?

US Healthcare Costs = $2.9 Trillion

17.4% of GDP

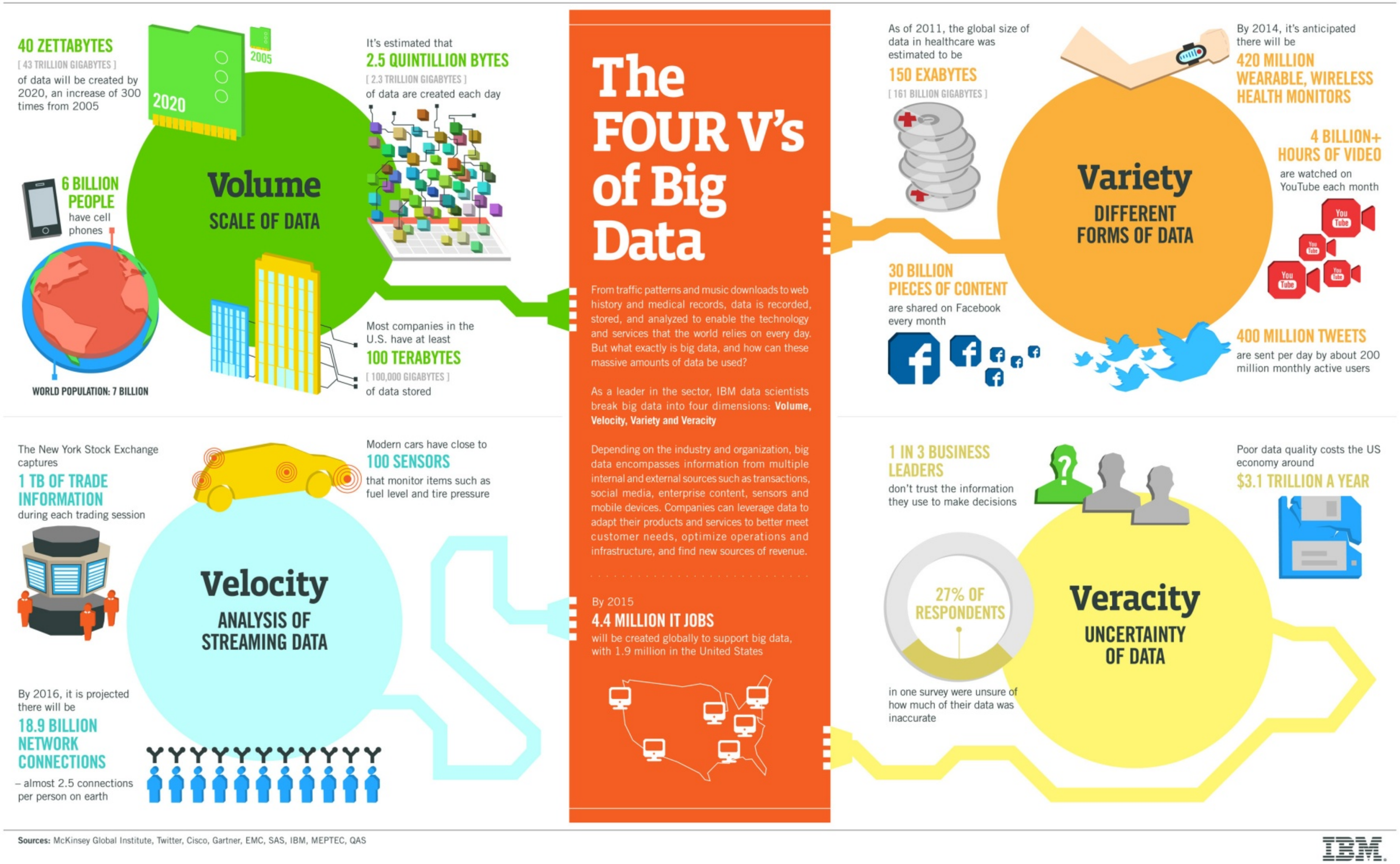
$9,255 per person

PROMISES	Savings of up to [illegible] a year. Shift toward "evidence-based" treatment.
OBSTACLES	Privacy concerns. Resistance from patients and doctors.

Big Data and Health Care

80% of all Health Care Data is unstructured

How does this Map to Health Care?



The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015 **4.4 MILLION IT JOBS** will be created globally to support big data, with 1.9 million in the United States

Volume
SCALE OF DATA

40 ZETTABYTES [43 TRILLION GIGABYTES] of data will be created by 2020, an increase of 300 times from 2005

It's estimated that **2.5 QUINTILLION BYTES** [2.3 TRILLION GIGABYTES] of data are created each day

6 BILLION PEOPLE have cell phones

WORLD POPULATION: 7 BILLION

Most companies in the U.S. have at least **100 TERABYTES** [100,000 GIGABYTES] of data stored

Variety
DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be **150 EXABYTES** [161 BILLION GIGABYTES]

By 2014, it's anticipated there will be **420 MILLION WEARABLE, WIRELESS HEALTH MONITORS**

4 BILLION+ HOURS OF VIDEO are watched on YouTube each month

30 BILLION PIECES OF CONTENT are shared on Facebook every month

400 MILLION TWEETS are sent per day by about 200 million monthly active users

Velocity
ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures **1 TB OF TRADE INFORMATION** during each trading session

Modern cars have close to **100 SENSORS** that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be **18.9 BILLION NETWORK CONNECTIONS** – almost 2.5 connections per person on earth

Veracity
UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS don't trust the information they use to make decisions

Poor data quality costs the US economy around **$3.1 TRILLION A YEAR**

27% OF RESPONDENTS in one survey were unsure of how much of their data was inaccurate

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

How does this Map to Health Care

- Volume (Data) & Scalability
  - Processing data where it resides. Locality of Reference - Genetics & Imaging & Brain Data sets. Bringing computation to the data.
- Variety (Data) & Extensibility – Imaging & EEG/MEG
  - Structured
  - Semi-structured
  - Unstructured
- Velocity (Data) & Computational Capability – Medical Wearables
- Veracity (Data) & Data Integrity & Semantics & Data – Clinical Trials
- Value (Data) - Hypothesis Generation

Other Health Care – Sharing & Collaboration

- Availability, Durability, Redundancy, Recoverability, Survivability, Longevity - Cloud
- Security (Systems), Privacy (Persons), de-identification (Protected Health Information) and Anonymity (Patient Records) – Global Controls
- Identity and Access Management – Global Controls
- Global Health Care Compliance
- Raw Data, Meta Data, Annotated and fully Curated Data and/or Analytical Results – genomic variation data.
- Seamless Integration – incorporate reference data sets, autonomous databases, health administration data, etc.
- Interoperability, Collaboration & Transparency

How big is the human genome?

In megabytes, not base pairs.

How big is the human genome?

- **In a perfect world (just your 3 billion letters): ~700 megabytes**

AGCCCCTCAGGAGTCCGGCCACATGGAAACTCCTCATTCCGGAGGTCAGTGATTTACCCTGGCTCACCTTGGCGTCGCGTCCGGCGGCAAACTAAGAACACGTCGTCTAAATGACTTCTTAAAGTAGAATAGCGTGTTCTCTCCTTCCAGCCTCCGAAAAACTCGGACCAAAGATCAGGCTTGTCCGTTCTTCGCTAGTGATGAGACTGCGCCTCTGTTCGTACAACCAATTTAGGTGAGTTCAAACTTCAGGGTCCAGAGGCTGATAATCTACTTACCCAAACATAG

How big is the human genome?

- **As a variant file, with just the list of mutations: ~125 megabytes**

Only about 0.1% of the genome is different among individuals, which equates to about 3 million variants (aka mutations) in the average human genome. This means we can make a “diff file” of just the places where any given individual differs from the normal “reference” genome. In practice, this is usually done in a .VCF file format, which in its simplest format looks something like so:

chr20 14370 rs6054257 G A 29 PASS 0|0

How big is the human genome?

- **In the real world, right off the genome sequencer: ~200 gigabytes (**FASTQ file format)

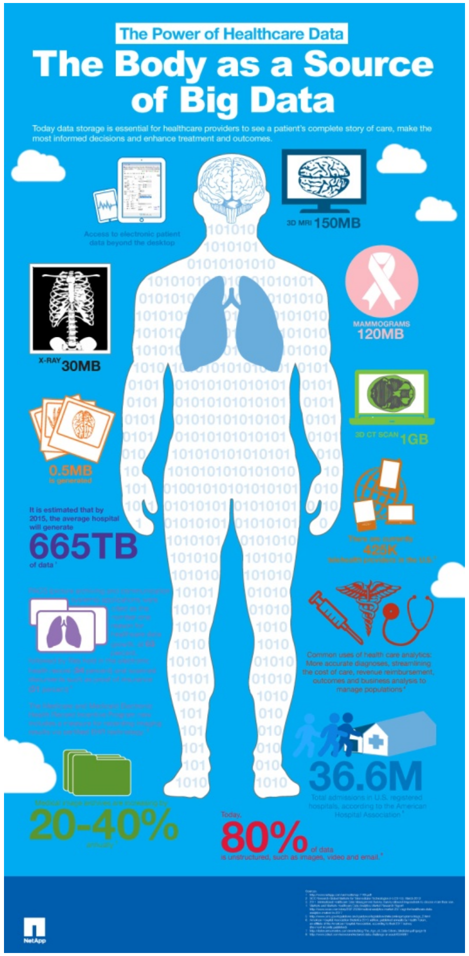
```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTG
TTCAACTCACAGTTT
+
!’’*((((***+))%%%++)(%%%%).1***-
+*’’))**55CCF>>>>>>CCCCCCC65
```

But this is how genomes are usually stored, because sequencing is still an imperfect science “So you really need to hang on to the raw sequencing reads and associated quality data, for future tweaking of the data analysis parameters if needed”.

How big is the human genome?

- What this means is that we’d all better brace ourselves for a major flood of genomic data.
- The 1000 genomes project data, for example, is now available in the AWS cloud and consists of >200 terabytes for the 1700 participants.
- As the cost of whole genome sequencing continues to drop, bigger and bigger sequencing studies are being rolled out. Just think about the storage requirements of this 10K Autism Genome project, or the UK’s 100k Genome project….. or even.. gasp.. this Million Human Genomes project.
- The computational demands are staggering.

Data Explosion in Medical Imaging



The Power of Healthcare Data

The Body as a Source of Big Data

Today data storage is essential for healthcare providers to see a patient's complete story of care, make the most informed decisions and enhance treatment and outcomes.

It is estimated that by 2015, the average hospital will generate 665TB of data[1]

There are currently 425K telehealth providers in the U.S.[2]

Common uses of health care analytics: More accurate diagnoses, streamlining the cost of care, revenue reimbursement, outcomes and business analysis to manage populations[4]

36.6M Total admissions in U.S. registered hospitals, according to the American Hospital Association[6]

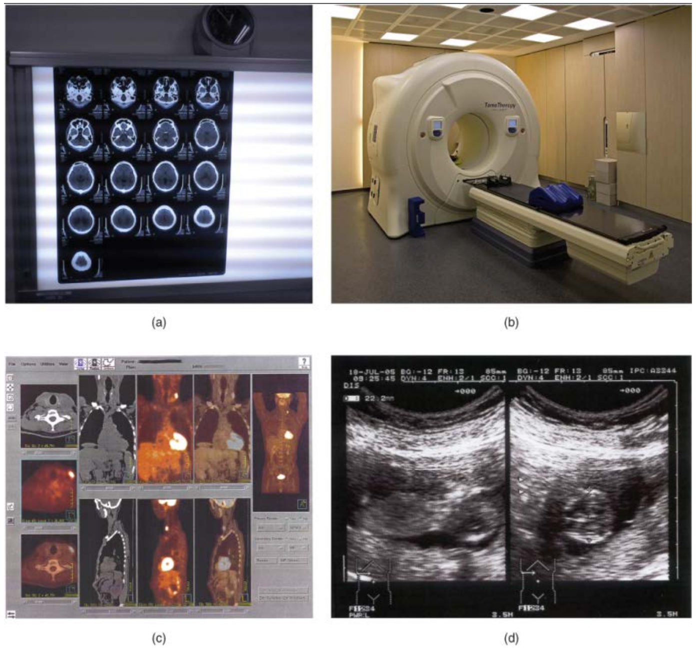
Medical image archives are increasing by 20-40% annually[7]

Today, 80% of data is unstructured, such as images, video and email.[8]

Data Explosion in Medical Imaging

- Medical Imaging modalities
  - 1.1Radiography
  - 1.2Magnetic Resonance Imaging (MRI)
  - 1.3Nuclear medicine
  - 1.4Ultrasound
  - 1.5Elastography
  - 1.6Tactile imaging
  - 1.7Photoacoustic imaging
  - 1.8Thermography
  - 1.9Tomography
    - 1.9.1Conventional tomography
    - 1.9.2Computer-assisted tomography
  - 1.10Echocardiography
  - 1.11Functional near-infrared spectroscopy

- Medical imaging topics
  - 2.1Image Gently and Image Wisely Campaigns
  - 2.2Maximizing imaging procedure use
  - 2.3Creation of three-dimensional images
  - 2.4Compression of medical images
  - 2.5Non-diagnostic imaging
  - 2.6Archiving and recording
  - 2.7Medical Imaging in the Cloud
  - 2.8Use in pharmaceutical clinical trials
  - 2.9Shielding

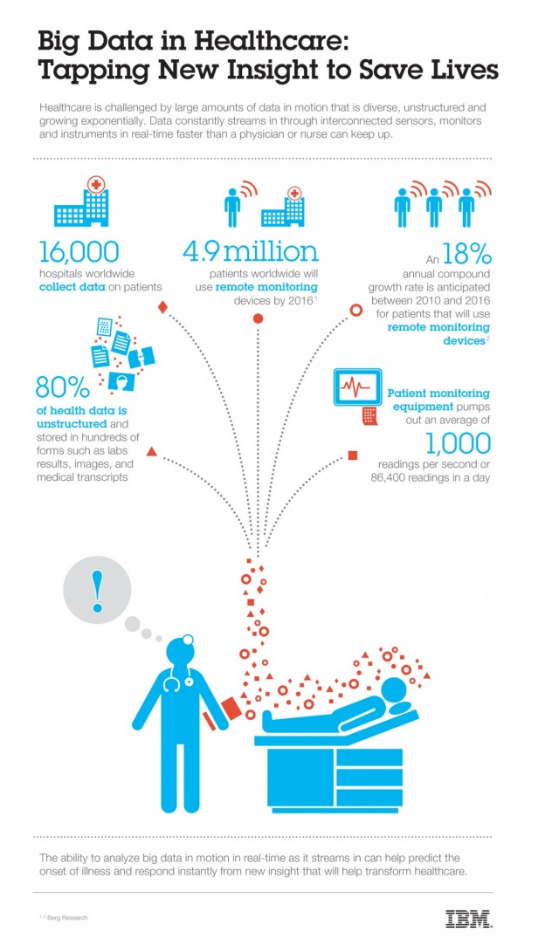


a) The results of a CT scan of the head are shown as successive transverse sections. (b) An MRI machine generates a magnetic field around a patient. (c) PET scans use radiopharmaceuticals to create images of active blood flow and physiologic activity of the organ or organs being targeted. (d) Ultrasound technology is used to monitor pregnancies because it is the least invasive of imaging techniques and uses no electromagnetic radiation.[3]

- As a discipline and in its widest sense, it is part of biological imaging and incorporates radiology which uses the imaging technologies of X-ray radiography, magnetic resonance imaging, medical ultrasonography or ultrasound, endoscopy, elastography, tactile imaging, thermography, medical photography and nuclear medicine functional imaging techniques as positron emission tomography (PET) and Single-photon emission computed tomography (SPECT).

- Measurement and recording techniques which are not primarily designed to produce images, such as electroencephalography (EEG), magnetoencephalography (MEG), electrocardiography (ECG), and others represent other technologies which produce data susceptible to representation as a parameter graph *vs.* time or maps which contain data about the measurement locations. In a limited comparison these technologies can be considered as forms of medical imaging in another discipline.

Data Explosion in Medical Imaging



Big Data in Healthcare: Tapping New Insight to Save Lives

Healthcare is challenged by large amounts of data in motion that is diverse, unstructured and growing exponentially. Data constantly streams in through interconnected sensors, monitors and instruments in real-time faster than a physician or nurse can keep up.

16,000 hospitals worldwide **collect data** on patients

4.9 million patients worldwide will use **remote monitoring** devices by 2016[1]

An 18% annual compound growth rate is anticipated between 2010 and 2016 for patients that will use **remote monitoring devices**[2]

80% **of health data is unstructured** and stored in hundreds of forms such as labs results, images, and medical transcripts

Patient monitoring equipment pumps out an average of 1,000 readings per second or 86,400 readings in a day

The ability to analyze big data in motion in real-time as it streams in can help predict the onset of illness and respond instantly from new insight that will help transform healthcare.

[1, 2] Berg Research

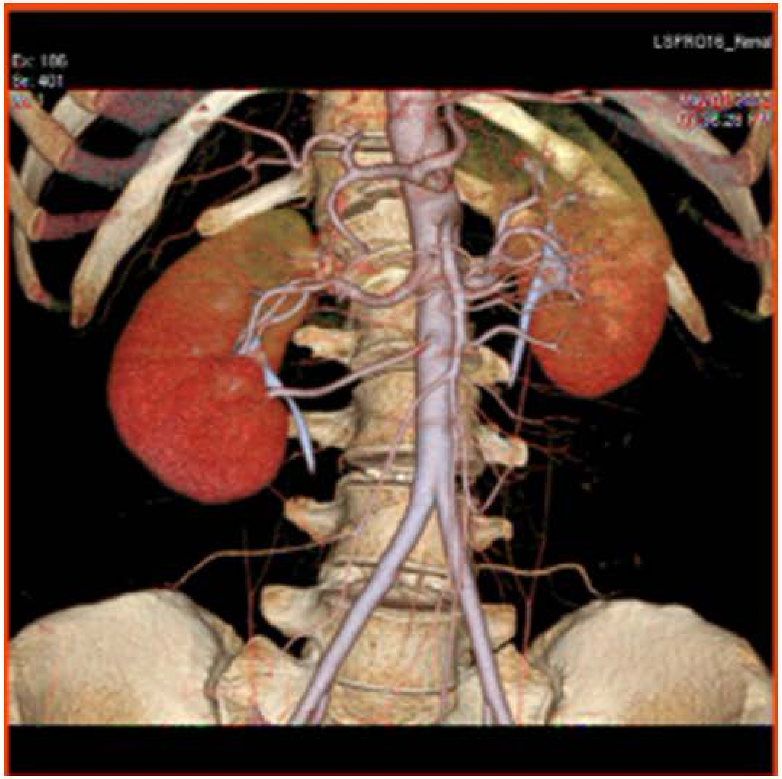
Data Explosion in Medical Imaging

- Wikipedia contributors. Medical imaging. Wikipedia, The Free Encyclopedia. January 17, 2016, 20:58 UTC. Available at: https://en.wikipedia.org/w/index.php?title=Medical_imaging&oldid=700325469. Accessed January 27, 2016.
- Data Explosion in Medical Imaging http://www.slideshare.net/sarcar/data-explosion-in-medical-imaging
- Pianykh, Oleg S. *Digital Imaging and Communications in Medicine (DICOM) A Practical Introduction and Survival Guide*, 2nd ed. Berlin: Springer, 2008. "A Practical Introduction and Survival Guide“
- Wikipedia contributors, "FASTQ format," *Wikipedia, The Free Encyclopedia,* https://en.wikipedia.org/w/index.php?title=FASTQ_format&oldid=686863247 (accessed January 27, 2016).
- Annotation and Image Markup – AIM, NIH, National Cancer Institute, NCI Wiki Created by Ann Wiley, last modified by Carolyn Klinger on Mar 24, 2015 https://wiki.nci.nih.gov/display/AIM/Annotation+and+Image+Markup+-+AIM

Digital X-Ray

Computed Tomography

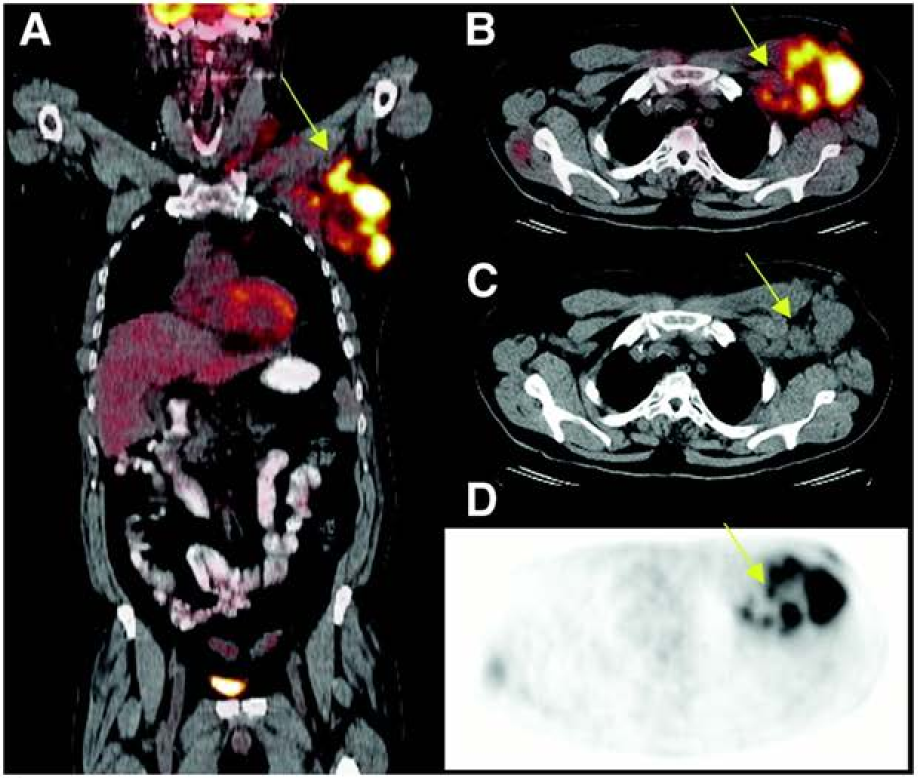
X-Ray technology, but in 3D !



Magnetic Resonance Imaging

- Principal of magnetic spin
- Great for soft tissue imaging

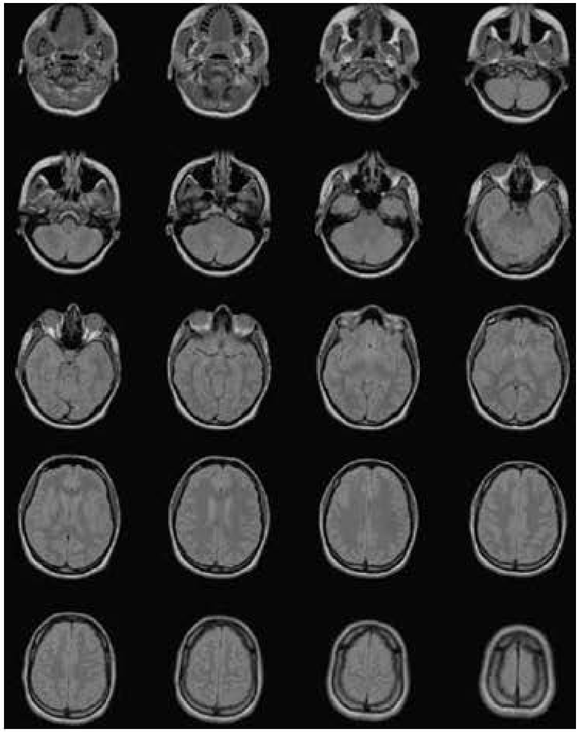
Positron Emission Tomography



PET-CT

- Functional imaging
- Radioactive bio-marker binds to *cancerous* cell
- Capture positron decay with a scintillation detector

How BIG is that Image Data ?



	Full body PET	CT Cardiac	fMRI
Images / set	600	3000	20000
Size of 1 set	1.2 GB	6 GB	40 GB
No. of sets (typical)	4	6	8
Exam Size	**9 GB**	**36 GB**	**300 GB**

Sizes are approximations

How BIG is that Medical Data?

Christian Medical College Vellore 0.5 million exams / yr	60 TB
Clalit Healthcare Services, 14-hospital network in Israel 4.5 million exams/yr	250 TB (annually)
Est. imaging data size in US – 2014	100 PB
Est. imaging data size globally – 2020	35 ZB

Technology

- Compression (lossless), Encryption
- Indexing and Searching of Data
- Parallel Everything
  - Files Systems
- Memory Residency
- Today's Multi-core, Multithreaded, Shared Nothing
- Open Source & Standards at all Health Care, Biomedical and Technology Levels
  - Standard Stacks – Google Genome
- Cloud
  - Federal Cloud Standard
- Data Management (NoSQL) – Integration at the analytics platform (SAS, SPSS, MathWorks, Cognos, etc.)
- Visualization Tools

Examples

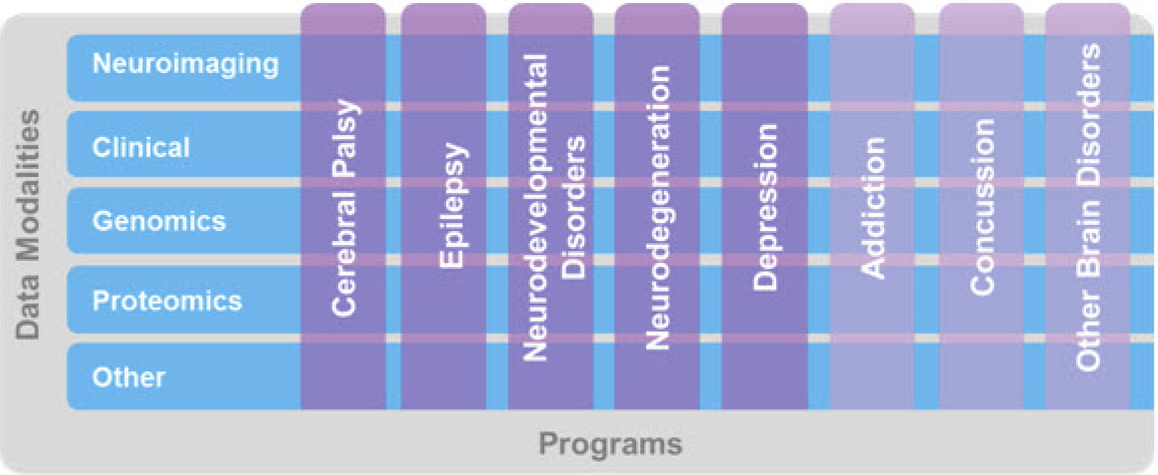


Figure 1. Multimodal data from multiple brain disorders are housed in Brain-CODE.

- The types of data available in Brain-CODE include clinical research data, Neuroimaging data (MRI, EEG, MEG, DTI), genomic data, proteomic data, and demographic data

The **Digital Imaging and Communications in Medicine** (DICOM) standard

http://dicom.nema.org

What is DICOM?

•DICOM is a global Information-Technology standard that is used in virtually all hospitals worldwide. Its current structure, is designed to ensure the interoperability of systems used to: Produce, Store, Display, Process, Send, Retrieve, Query or Print medical images and derived structured documents as well as to manage related workflow.

•DICOM is used to aid the distribution and viewing of medical images, such as CT scans, MRIs, and ultrasound.

•DICOM is used in: · radiology · breast imaging · cardiology · radiotherapy · oncology · ophthalmology · dentistry · pathology · surgery · veterinary · neurology

- http://dicom.nema.org/dicom/geninfo/Brochure.pdf It was created by the **National Electrical Manufacturers Association** (**NEMA**) to aid the distribution and viewing of medical images, such as CT scans, MRIs, and ultrasound. **Part 10 of the standard describes a file format for the distribution of images.**

DICOM Standard

20
parts

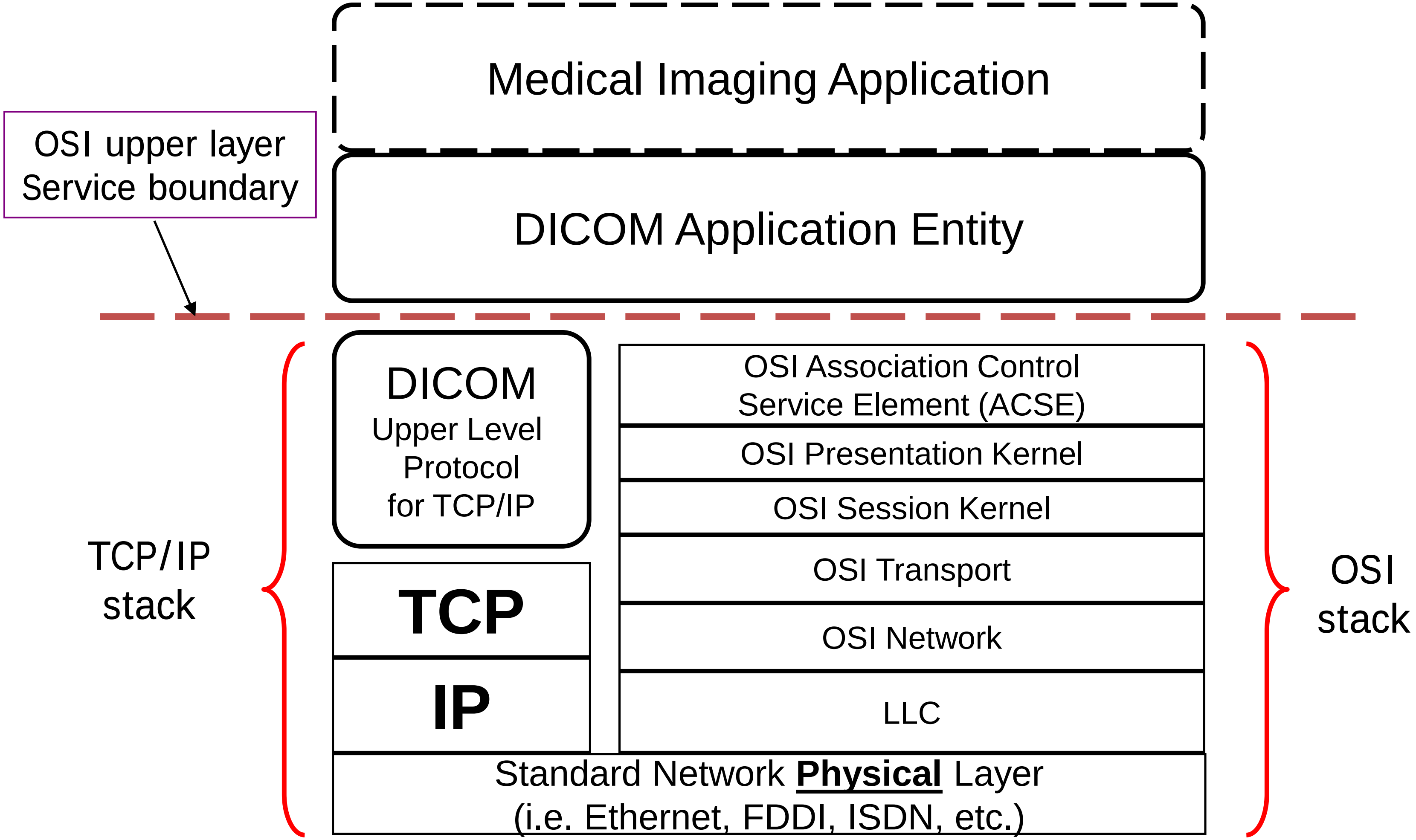
+

161
supplements

PS 3.1: Introduction and Overview
PS 3.2: Conformance
PS 3.3: Information Object Definitions
PS 3.4: Service Class Specifications
PS 3.5: Data Structure and Encoding
PS 3.6: Data Dictionary
PS 3.7: Message Exchange
PS 3.8: Network Communication Support for Message Exchange
PS 3.9: Point-to-Point Communication Support for Message Exchange (Retired)
PS 3.10: Media Storage and File Format for Data Interchange
PS 3.11: Media Storage Application Profiles
PS 3.12: Storage Functions and Media Formats for Data Interchange
PS 3.13: Print Management Point-to-Point Communication Support (Retired)
PS 3.14: Grayscale Standard Display Function
PS 3.15: Security Profiles
PS 3.16: Content Mapping Resource

- PS 3.17: Explanatory Information
- PS 3.18: Web Access to DICOM Persistent Objects
- PS 3.19: Application Hosting
- PS 3.20: Transformation of DICOM to and from HL7 standards

DICOM Network: where is it in the network stack?



Medical Imaging Application

DICOM Application Entity

OSI upper layer Service boundary

TCP/IP stack

DICOM Upper Level Protocol for TCP/IP

TCP

IP

OSI Association Control Service Element (ACSE)

OSI Presentation Kernel

OSI Session Kernel

OSI Transport

OSI Network

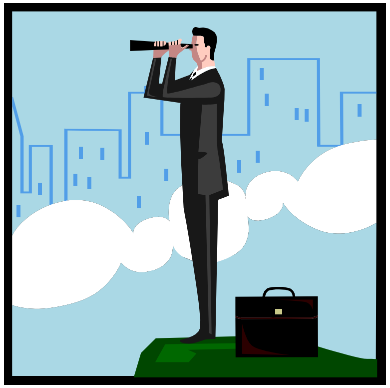
LLC

Standard Network **Physical** Layer (i.e. Ethernet, FDDI, ISDN, etc.)

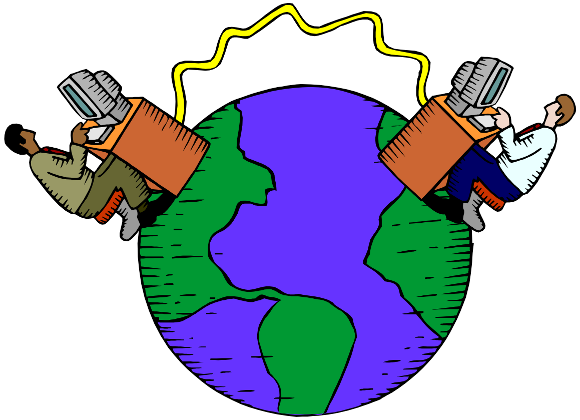
OSI stack

Challenges of Large Imaging Data

Archival



Search



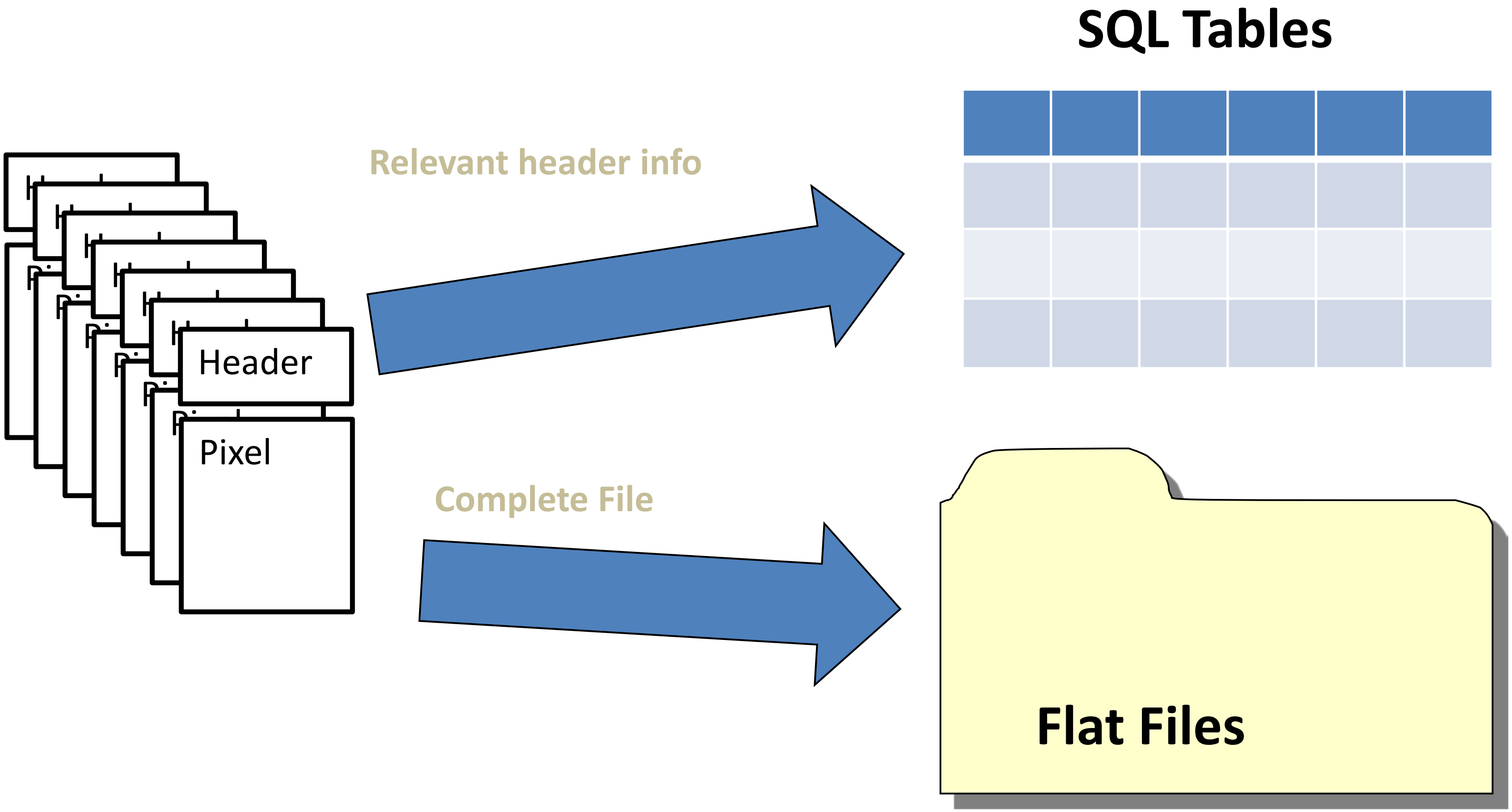
Transfer

Lawmakers demand storage guarantee

Moms:
“25 years after the birth of the last child”

Mentally disabled:
“20 years after the last contact or 8 years after the patient's death”

Children:
“Until the patient is 25”



SQL Tables

Relevant header info

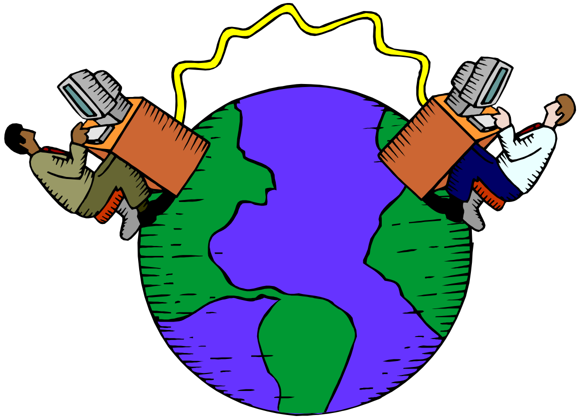
Header

Pixel

Complete File

Flat Files

Challenges of DICOM



- Decentralized Storage
- DICOM is based on TCP/IP
  - Slow over large number of hops
  - CISCO WAAS
- DICOM compression is not adequate
  - Lossy, Loseless
- DICOM is not efficient on fault-tolerance
  - Dated retry mechanism, transmit in sets/series, not files

FOSS DICOM Tools and Images

API

Language	Toolkit
C/C++	GDCM, DCMTK
Java	Pixel, dmc4che
Perl	DICOM.pm
Ruby	Ruby DICOM
Python	pydicom
PHP	Nanodicom
C#	DICOM#

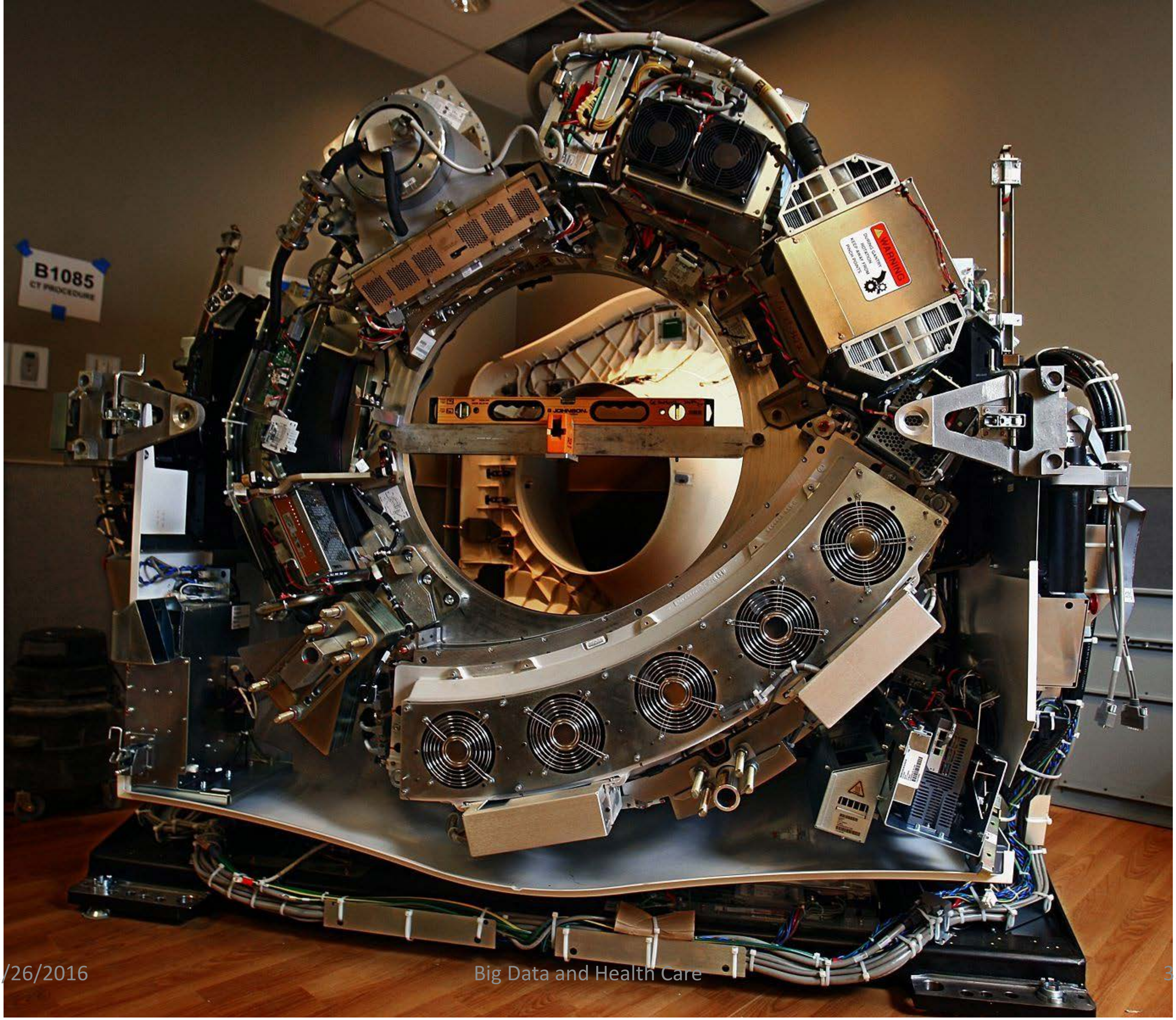
Viewers

OsiriX for Mac
Santesoft for Win
Kradview for Linux

Public Datasets

ftp://medical.nema.org/medical/dicom/DataSets/

http://www.barre.nom.fr/medical/samples/


B1085
CT PROCEDURE
WARNING
JOHNSON

Annotation and Image Markup - AIM NIH – National Cancer Institute

AIM is the first project to propose and create a standard means of adding information and knowledge to an image in a clinical environment, so that image content can be easily and automatically searched. AIM provides a solution to the following imaging challenges:

- No agreed upon syntax for annotation and markup
- No agreed upon semantics to describe annotations
- No standard format (for example, DICOM, XML, HL7) for annotations and markup

The AIM project includes the following tools.

- The AIM Model captures the descriptive information for an image with user-generated graphical symbols **placed on the image** into a single common information source.

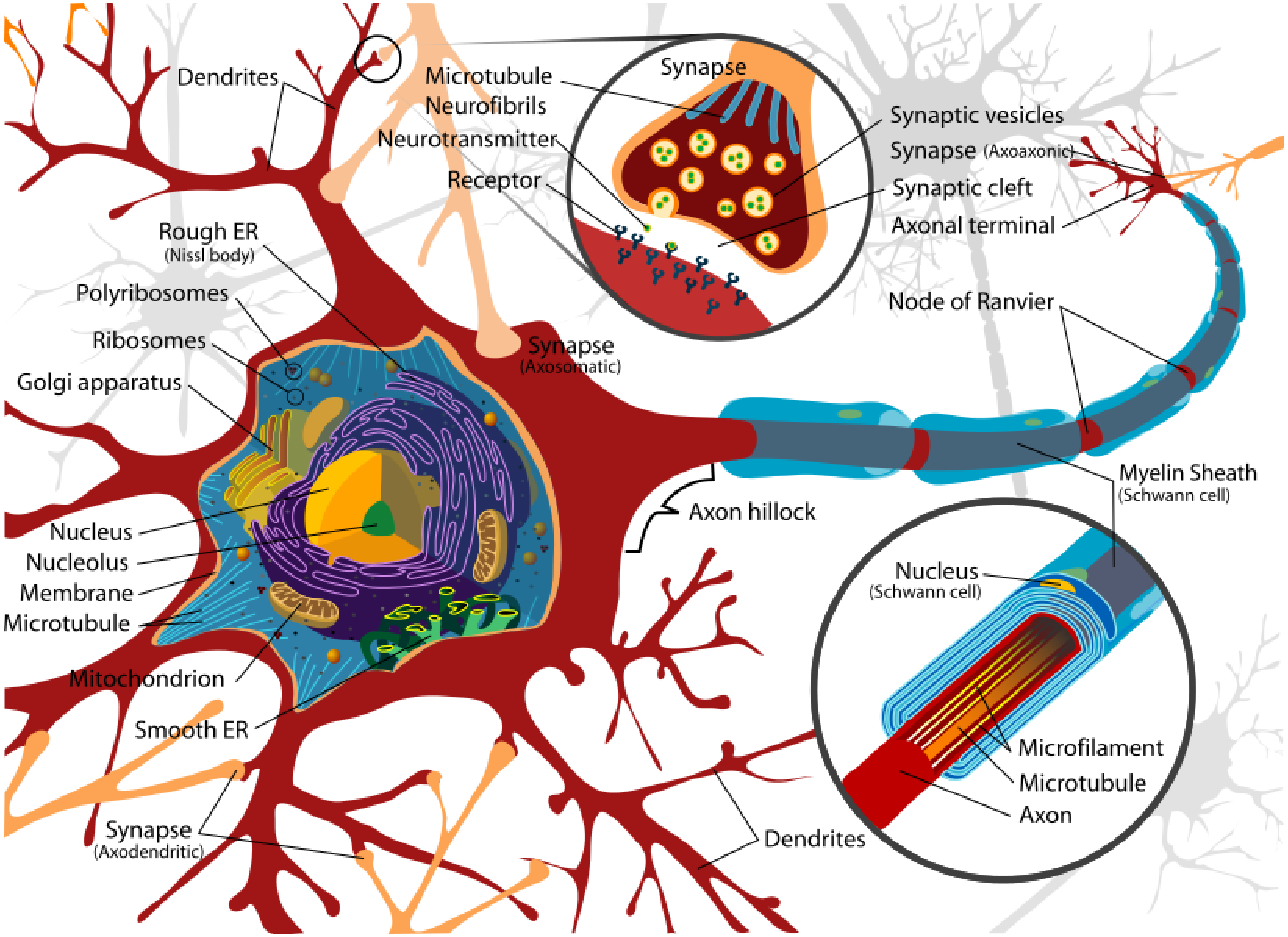
Data Explosion in Medical Imaging

Harvard Institute for Applied Computational Science

FIFTH ANNUAL SYMPOSIUM ON THE FUTURE OF COMPUTATION IN SCIENCE AND ENGINEERING - BRAIN + MACHINE

- https://www.youtube.com/watch?v=v3O_avy0uys
- https://www.youtube.com/watch?v=DAfZy2K6Njk&list=PLfjZYvoyxDtZjm9wJVxQIOU8SvoUBG3sQ

Next Stop – The Human Brain



How many neurons & synapses make a human brain?

- There are approximately **86 billion** (**86,000,000,000**) neurons in the human brain.
- A typical neuron fires 5 - 50 times every second. Each individual neuron can form thousands of links with other neurons in this way, giving a typical brain well over **100 trillion synapses** (up to 1,000 trillion, by some estimates).