

Practical Applications and Pitfalls Of 'Big Data' For Decision Support In Medical Imaging and Informatics

Potential Collaborations with Medical Imaging and NASA

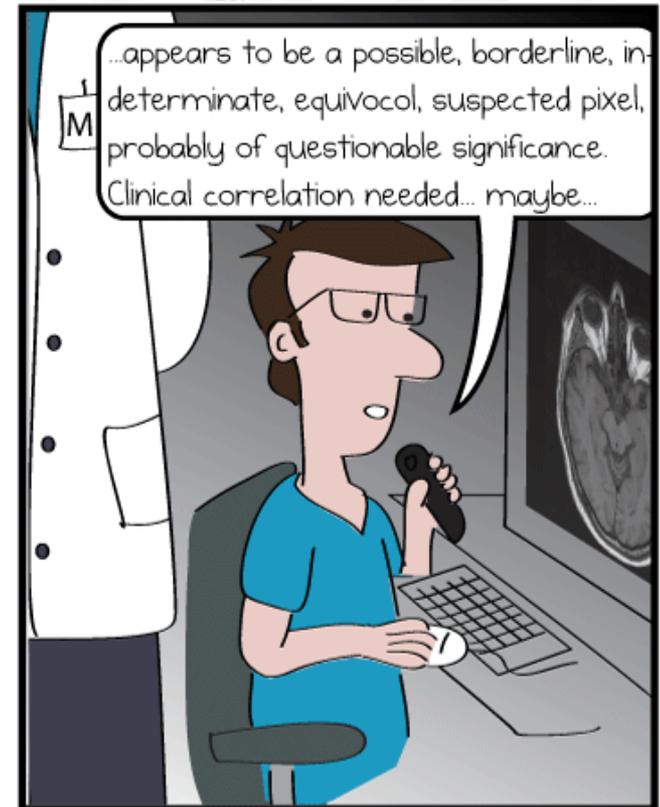
Eliot Siegel, MD, FACR, FSIIM

Professor and Vice Chair University of Maryland School of Medicine
Department of Diagnostic Radiology

Professor Computer Science University of Maryland Baltimore County

Professor Biomedical Engineering University of Maryland College Park

- What does a radiologist do?
- 8 year old perspective



NASA Explores Space Diagnostic Radiology and Nuclear Medicine Explore Inner Space





Diagnostic Imaging Decision Support Time Warp Since 1993

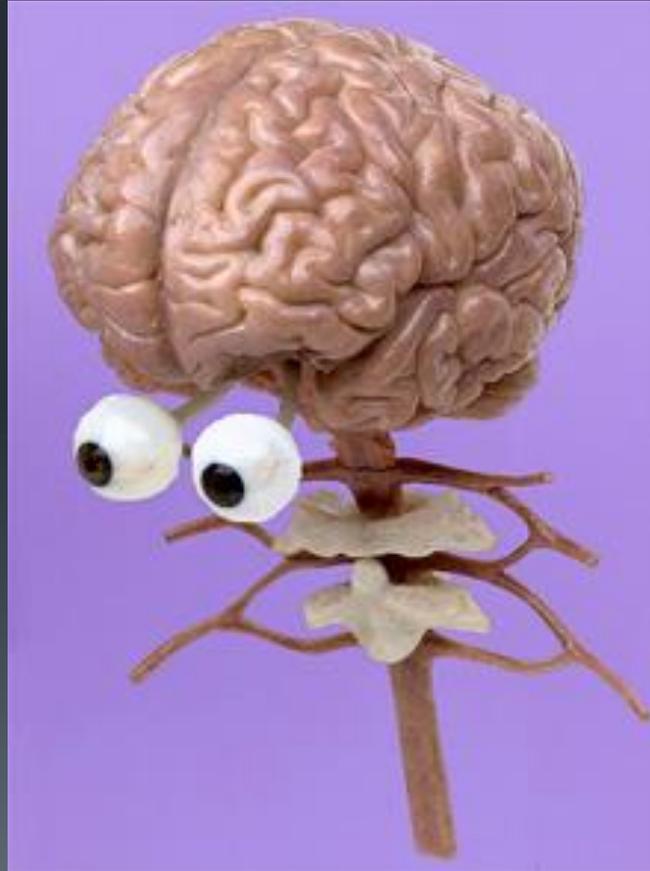
- We expected elimination of “lost films” and improved productivity using cine/stack mode and rapid display and navigation of images but we had hoped for so much more!
- 1993 was around the time of a series of exciting papers predicting the imminent arrival of computer aided detection/decision support in radiology

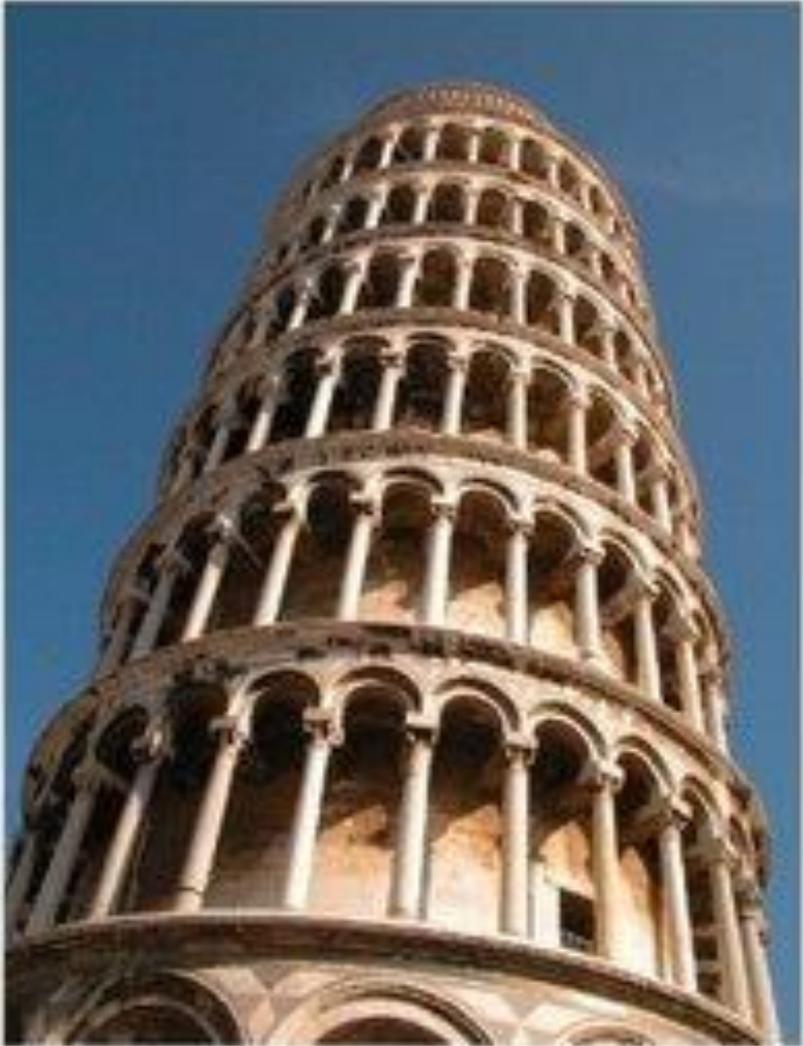
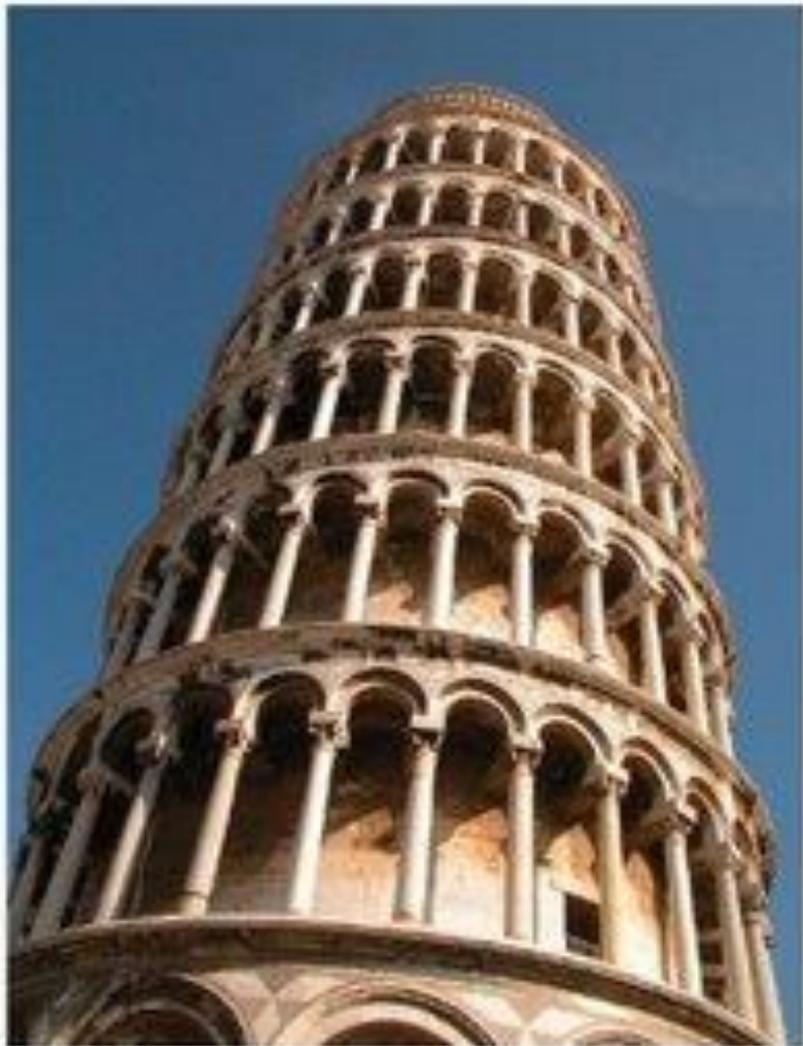
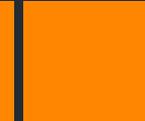
I believe this Lack of Use of Computers for Diagnostic and Therapeutic Decision Support Current Anecdotal Practice of Medicine and Diagnostic Imaging will Make it Seem Like 2015 was Flintstone Era in 10 to 20 Years



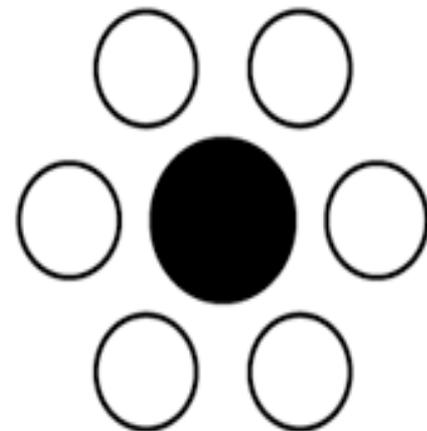
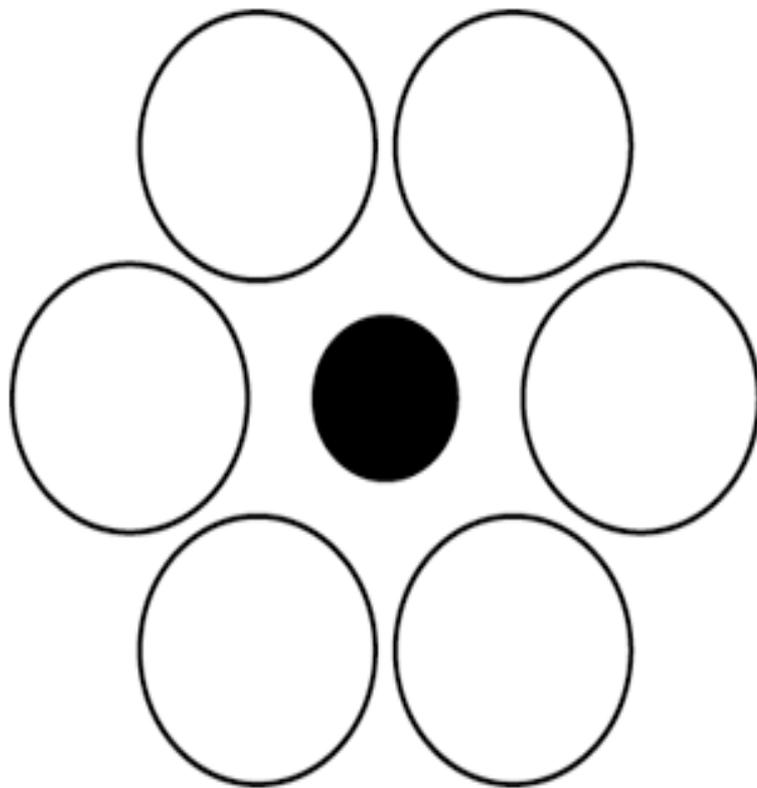
Do We Really Need Computer Assistance?

How Much Can We Rely on Our Human Eyes and Brain?

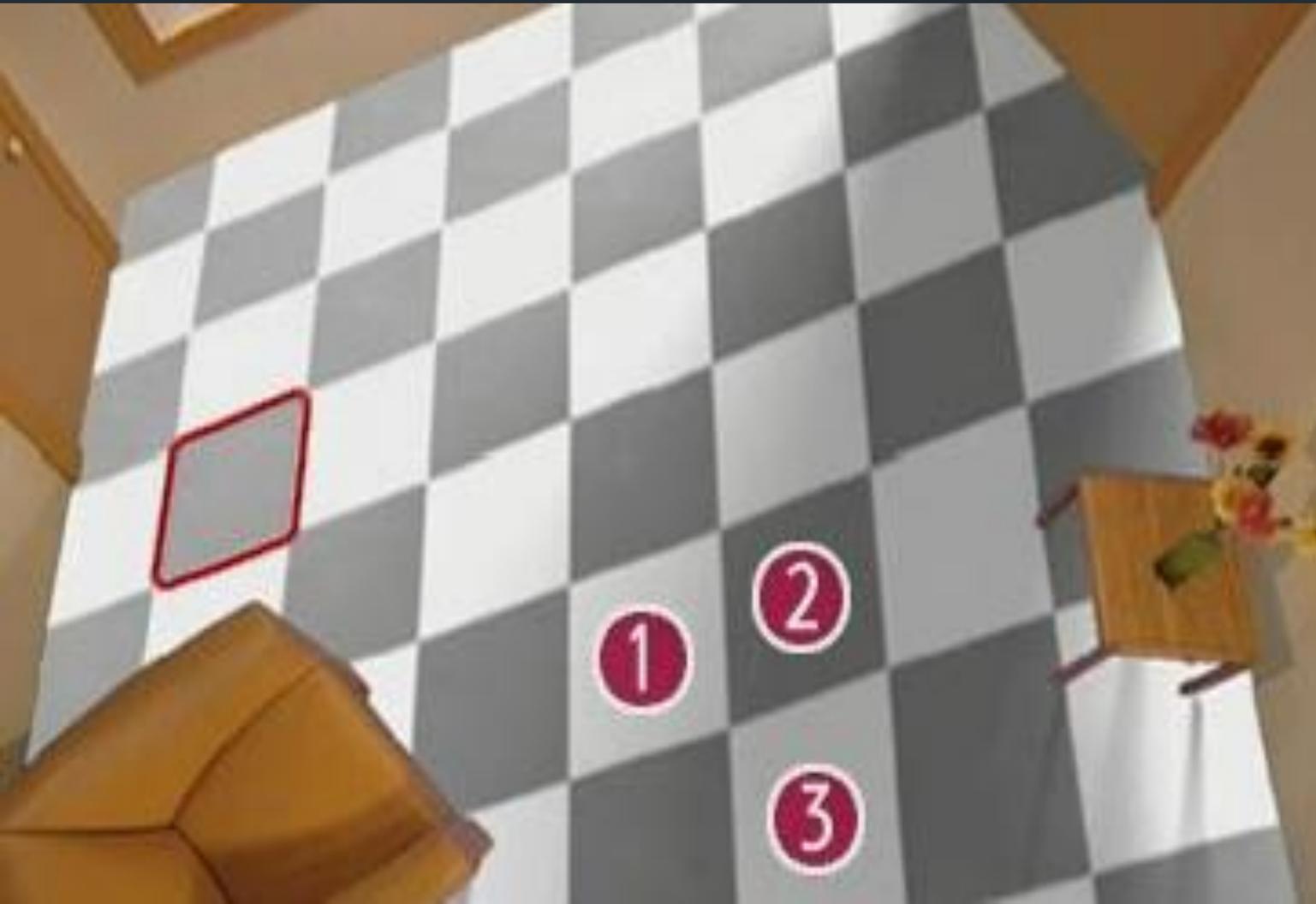


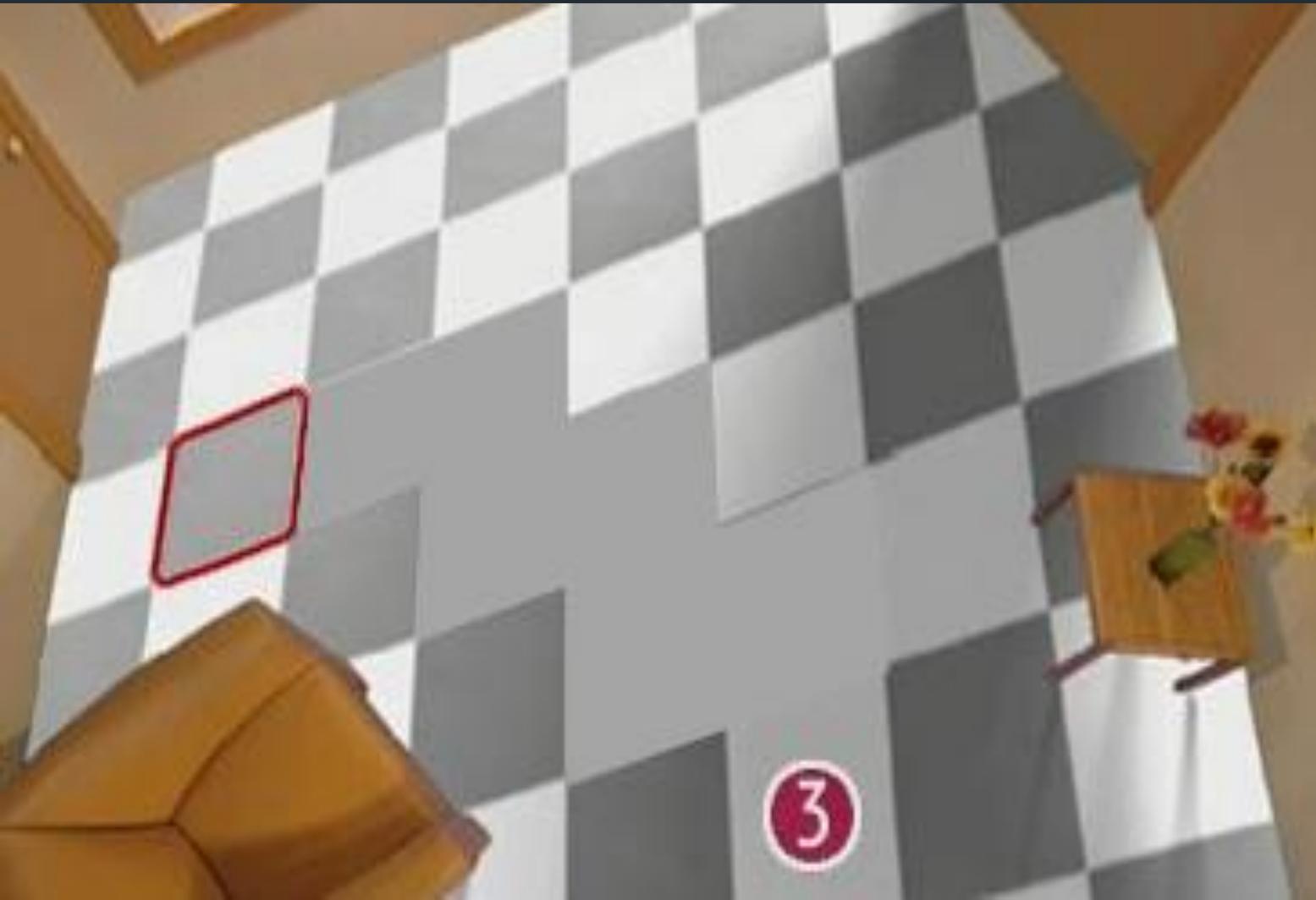


Which Black Ball is Bigger?

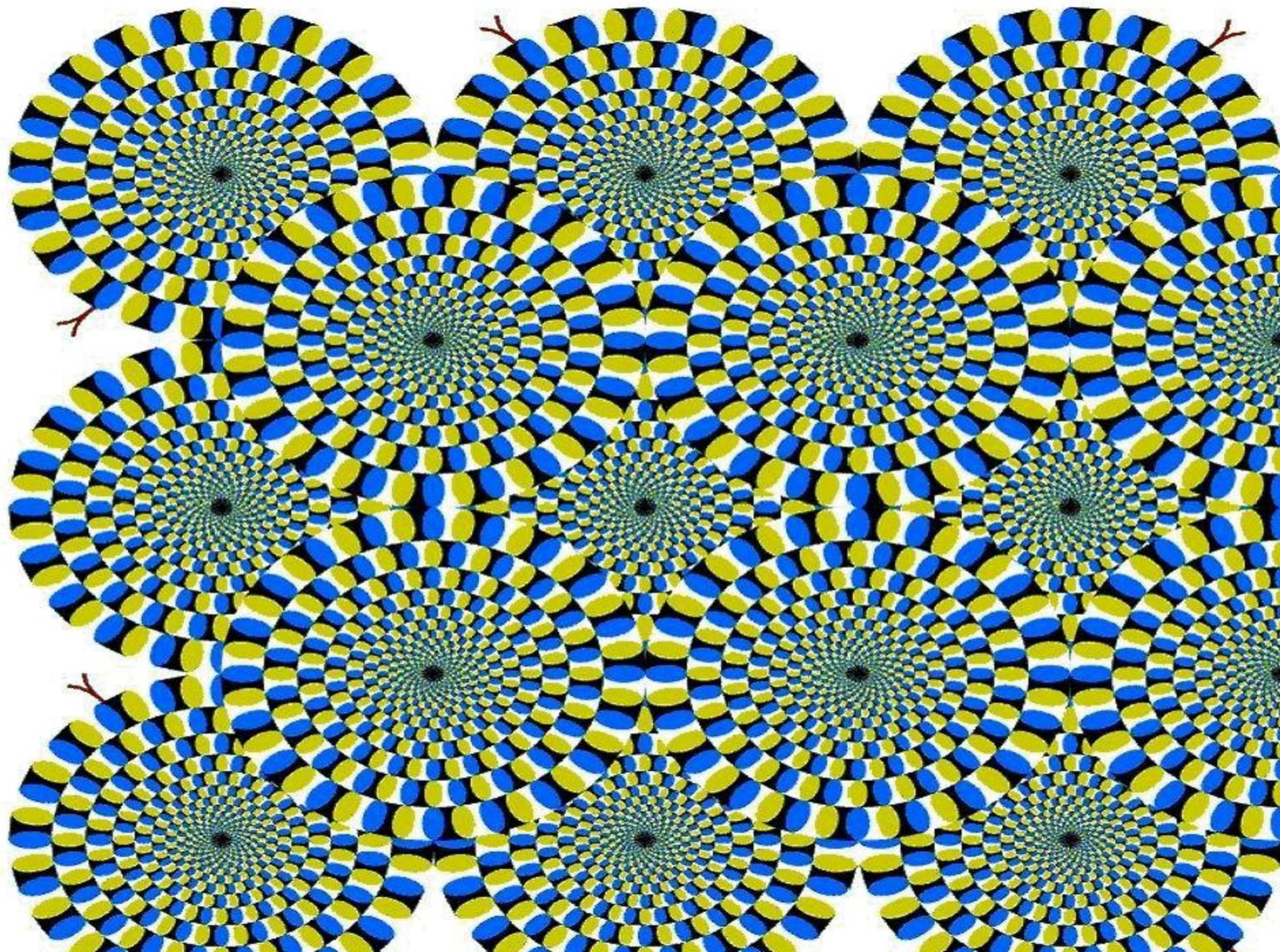


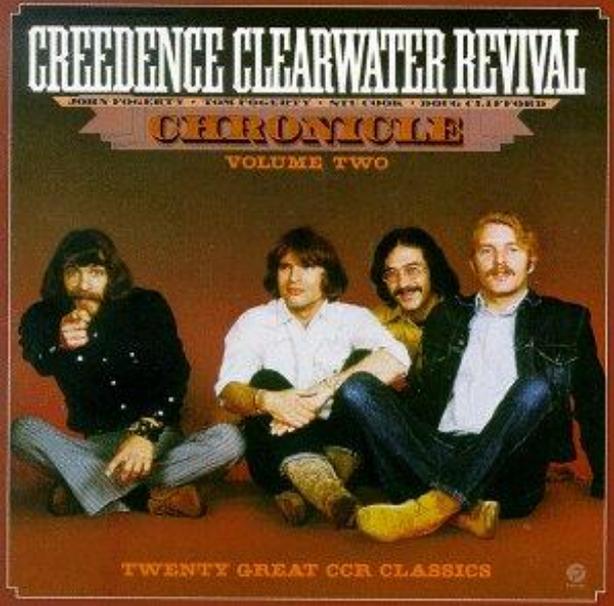
Which Square(s) Match The One Outlined in Red?





3

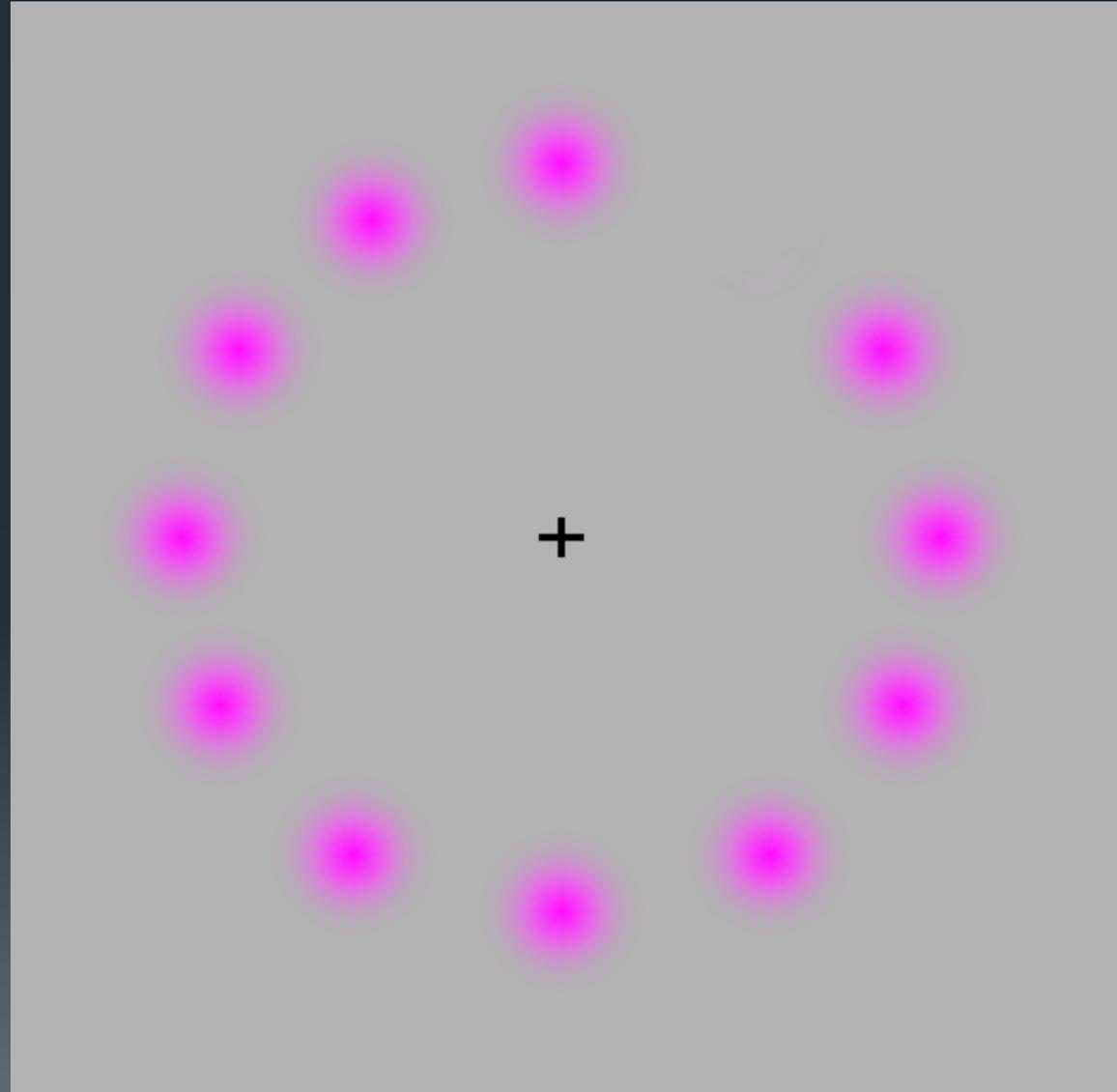




Creedence Clearwater Revival and Marvin Gaye “I Heard It Through the Grapevine”



- ... “People say believe half of what you see,
Son, and none of what you hear.
I can't help bein' confused” ...



Spinning Ferris Wheel



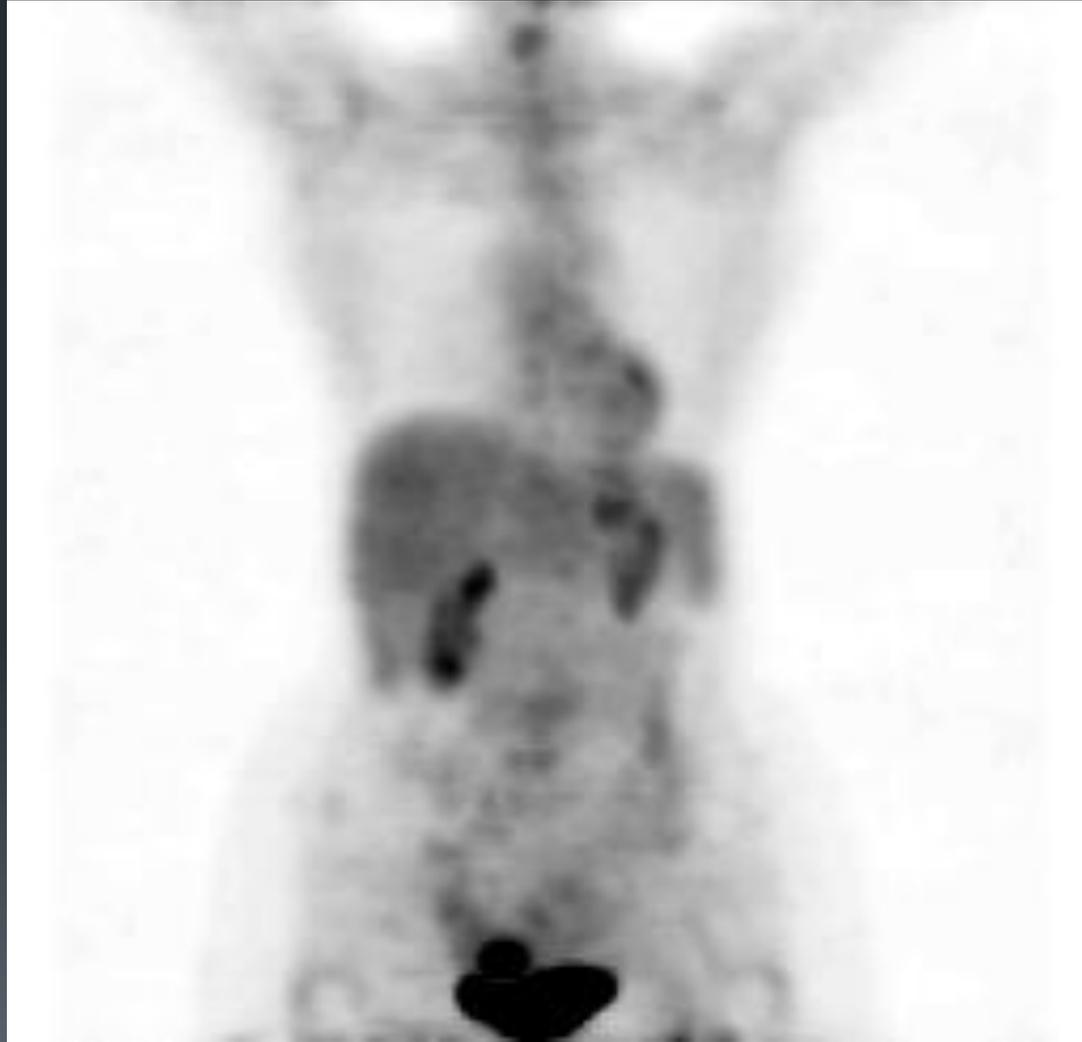


Which Direction is
the Dancer

Spinning
(clockwise or
counterclockwise)
and which foot is
she putting down
on the ground?

Raise your Hand
When She
Changes
Direction?

**Which Direction is this PET Scan Rotating?
Counterclockwise, clockwise, oscillating?**



Assistance is Improved Productivity

Job Switching – Lucy and Ethel at the Candy Factory



With Increased Diagnostic Tools (US, CT, MRI, PET) and Techniques and Discoveries in the Journals It Is Impossible to keep up with the Literature Especially in Rapidly Advancing Areas Such as Genomics

17



- ***The volume and complexity of medical information in healthcare continues to accelerate, recently doubling in less than five years with 80% or more of that data unstructured***



Is There A Need for Big Data and Analytics (or Artificial Intelligence) In Medicine?



Motivation for Artificial Intelligence Software in Medicine

- Schiff
 - Diagnostic errors far outnumber other medical errors by 2-4X
- Elstein
 - **Diagnostic error rate of about 15%** in line with autopsy studies
- Singh and Graber
 - Diagnostic errors are single largest contributor to ambulatory malpractice claims (40% in some studies) and cost about \$300,000 per claim
- Graber
 - Literature review of causes of diagnostic error suggest 65% system related (e.g. communication) and **75% had cognitive related factors**



Cognitive Errors

Graber et al Diagnostic Error in Internal Medicine,
Arch Intern Med 2005; 165:1493-1499

- Cognitive errors primary due to “faulty synthesis or flawed processing of the available information”
- Predominant cause of cognitive error was **premature closure** (satisfaction of search in diagnostic imaging)
 - Failure to continue considering reasonable alternatives after an initial diagnosis was reached



Cognitive Errors

- Other contributors to cognitive errors
 - Faulty context generation – lack of awareness of aspects of patient info relevant to diagnosis
 - Misjudging salience of a finding
 - Faulty detection or perception
 - Failed use of heuristics – assuming single rather than multifactorial cause of patient symptoms

Cognitive Errors

- Graber suggested augmenting “a clinician’s inherent metacognitive skills by using expert systems”
- The type of errors computers and humans make are different and thus the two working together are complementary
- Stated that clinicians continue to miss diagnostic information and “one likely contributing factor is the overwhelming volume of alerts, reminders, and other diagnostic information in the Electronic Health Record”



How Reliable is Our Judgment of Likelihood of Disease?

Ebola Screening

- Brand new extraordinarily high accuracy urine test just gets released The sensitivity of the test is truly impressive at 98% (If the subject has Ebola the test is positive 98% of the time)
- Specificity of the test is even higher, a “remarkable” 99% so if the subject does not have Ebola, the test comes back negative 99% of the time
- Say at the airport in Liberia 10 in every 50,000 people that fly have Ebola at any given time
- So if the test comes back positive what is the probability that the passenger actually has Ebola?

Choose the Best Answer: If a Passenger Tests Positive the Likelihood of Ebola will be:

- A. Greater than 99% given the combined sensitivity and specificity both greater than 98%
- B. Actually equal to the sensitivity of the test which is 98%
- C. Equal to the specificity of the test which is 99%
- D. Only approximately 50%
- E. Only approximately 8.6%
- F. Less than 2%

Ebola Screening Accuracy if Positive Result <2%! Because even with 99% specificity there are 500 false positives and ten true positives

- Actually 1.92% probability that a person has Ebola if the sensitivity is 98% and the specificity is 99%

$$\begin{aligned} P(A | B) &= \frac{P(A)P(B | A)}{P(A)P(B | A) + P(\bar{A})P(B | \bar{A})} \\ &= \frac{0.0002 \cdot 0.98}{0.0002 \cdot 0.98 + 0.9998 \cdot 0.01} \\ &= \frac{98}{5097} \\ &\approx 1.92\%. \end{aligned}$$

- In order to have the information necessary to more precisely determine the likelihood that a patient's images suggest a given diagnosis we need to:
 - Have good quality quantitative information
 - Good quality **a priori information** about the likelihood a given person has the disease outside the information in the imaging study
 - Critical importance of “personalized” clinical information on patient
 - But also critical importance of have data on large numbers of similar patients/similar images

Big Data Will Play a Critical Role in Helping in the Practice of Clinical Radiology?

What is Big Data?

- *Currently one of the hottest topics in medicine from both a research and clinical perspective*
- *Impossible to get any consensus on its definition and it seems to vary depending on one's perspective*
- *Later in this session, Dr. Lindsköld will share his definition of “big data” as that which “no longer fits into Excel!”*
- *The National Institute of Standards and Technology defines big data as **that which “exceed(s) the capacity or capability of current or conventional methods and systems***

40 ZETTABYTES

[43 TRILLION GIGABYTES]

of data will be created by 2020, an increase of 300 times from 2005

6 BILLION PEOPLE have cell phones



WORLD POPULATION: 7 BILLION

Volume SCALE OF DATA



It's estimated that 2.5 QUINTILLION BYTES

[2.3 TRILLION GIGABYTES]

of data are created each day



Most companies in the U.S. have at least

100 TERABYTES

[100,000 GIGABYTES] of data stored

The New York Stock Exchange captures

1 TB OF TRADE INFORMATION

during each trading session



Velocity ANALYSIS OF STREAMING DATA



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



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



As of 2011, the global size of data in healthcare was estimated to be

150 EXABYTES

[161 BILLION GIGABYTES]



30 BILLION PIECES OF CONTENT

are shared on Facebook every month

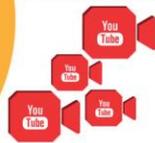


Variety DIFFERENT FORMS OF DATA



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



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



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

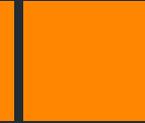


27% OF RESPONDENTS

in one survey were unsure of how much of their data was inaccurate

Veracity UNCERTAINTY OF DATA

- *Whichever of the above or other definitions that we may use when thinking of “Big Data”, medicine and in specific, **diagnostic imaging clearly generates vast amounts of it particularly “high dimensional” data***



Imaging: Like Claude Rains, The Invisible Man?

- *One of the major challenges with medical imaging is the difficulty of discovery of imaging information in the electronic medical record and from clinical trial data*



- 
- *Our imaging reports are, almost without exception, unstructured and our medical images are rarely tagged in such a way as to be discoverable or useful to data mining efforts*
 - *This must change if medical imaging is to play a substantial role in this era of big data, medical guidelines, decision support and personalized medicine.*

Personalized/Precision Medicine: Most Recently Closely Associated with Genomics

- The term “personalized medicine” or “precision medicine” came up in President Obama’s 2015 State of the Union Address
- It has recently been widely applied to describe the concept of providing medical care based on genetic differences between patients
- Access to genetic information will radically change the way medicine is practiced



Personalized Medicine



- In the near future, detailed information about each patient's unique genetic makeup will be available to their doctors
- It will be used in the selection of the best specific diagnostic and therapeutic methods for caring for that individual and for the more general purposes of lifestyle and/or **wellness counseling**

Learn valuable health & ancestry information.



- Reports on 240+ health conditions and traits
- Testing for 40+ inherited conditions
- Discover your ancestry composition
- Updates on your DNA as science advances

\$99

Order Now

One Price. Enjoy a subscription-free, ongoing service.

More features than anyone else.

\$99
SUBSCRIPTION FREE

Discover your:

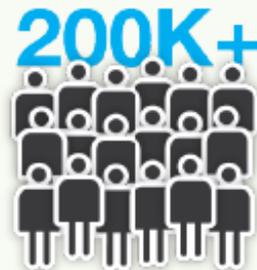
- + Ancestry composition
 - + Genetic relatives
 - + 23andMe Family Tree
 - + Maternal & paternal lineages*
 - + Neanderthal percentage
- + And much more!



**Paternal lineage is only scientifically possible for male members.*

Over 200,000 genotyped members.

We have the largest genealogical DNA database in the world. More matches, more data, more discoveries.



750,000

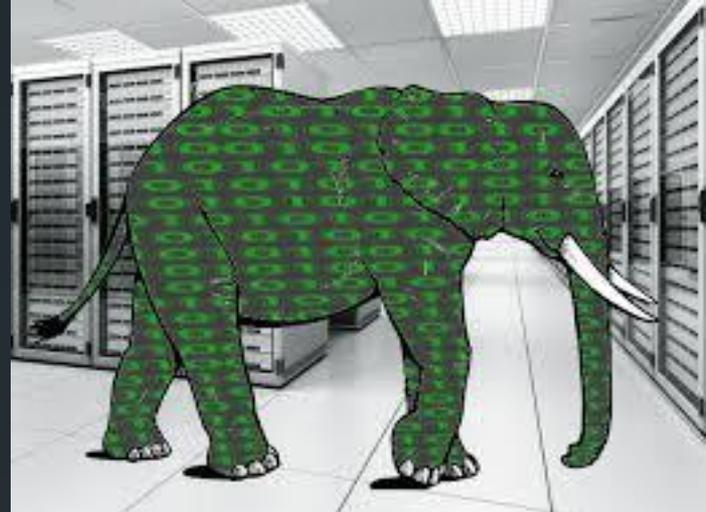
Recently featured on PBS' "Finding Your Roots."

23andMe genotyped the celebrities on "Finding Your Roots" to help them map their heritage and discover their global origins.

finding your roots
with HENRY LOUIS GATES, JR.



When Mining Big Data, Every Patient's Medical Care Becomes a Clinical Trial



- Rather than relying on a small study, or oncologist's personal experience, each patient's clinical course and data will be saved and made available for decision support rather than just the **2-3%** of oncology patients that are currently enrolled in clinical trials

If Medical Parameters Are Complex, Big Data, Imaging is Arguably Orders of Magnitude More So Compared with Even Genomic, Proteomic Data

- ***Imaging characterized by large number of studies with more sequences, images per study and large variety of different imaging procedures but this is not quite big data***
- **MRI**
 - T1, T2, Proton density, Echo planar, Inversion Recovery, Perfusion, Diffusion-Kurtosis, Spectroscopy, many contrast agents and many sequences
- **US**
 - Contrast, elastography, flow
- **CT**
 - Dynamic contrast
 - Multi-spectral

Molecular Imaging/PET Parameters

- Dozens radiopharmaceuticals
- PET
 - Glucose utilization: FDG,
 - Tumor Cellular Proliferation: ^{18}F -FLT-PET and tumor cellular proliferation
 - Tumor Hypoxia: [^{18}F]fluoromisonidazole (F-MISO)
 - Apoptosis: [^{18}F]ICMT-11
 - Additional PET
 - [^{18}F]fluoroethyl-l-tyrosine (FET)
 - [^{18}F]fluoro- α -methyltyrosine (FMT)
 - 6- [^{18}F]fluoro-dihydroxy-l-phenylalanine (F-DOPA)
 - [^{11}C]choline (CHO) and [^{18}F]choline.

How Big is Big Data in Imaging?



- Medical images are “large” and it’s been suggested that they represent more than 90% of the total storage space in healthcare systems but this is not what we mean by big data
- The number of imaging studies performed has increased precipitously but this is not what I think of as big data
- Unlike other healthcare data, medical images are extraordinarily complex and contain vast “hidden” amounts of information that go largely untapped



Radiology as Dark Matter

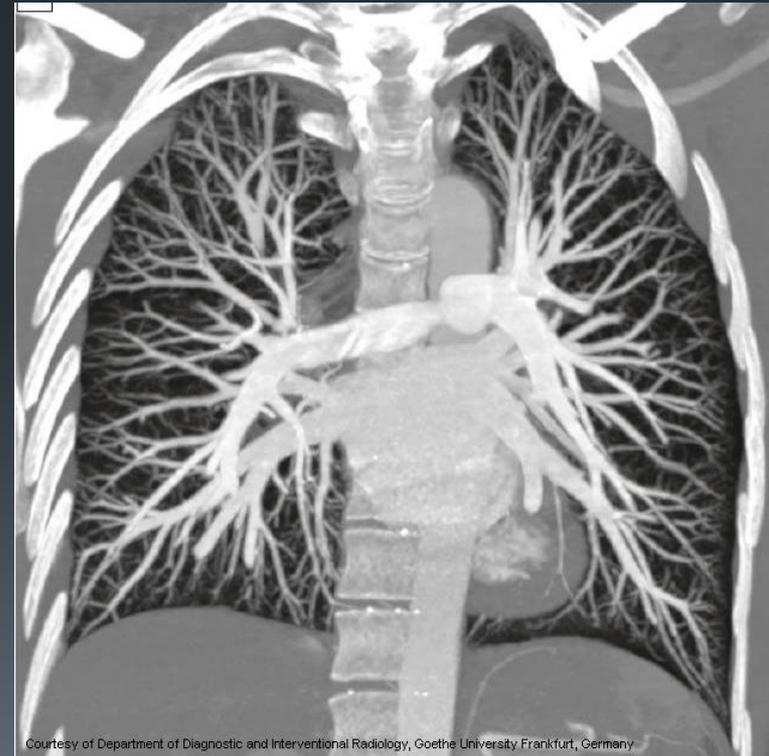
- As with Dark matter (27% of the entire universe)/Dark energy (68% of the universe), only 5% of the universe is observable with our current instruments
- This is similar to the case in radiology where even a smaller percentage of our image “data” is in our interpretation and the vast majority locked up within the images themselves

Why Do You Keep Your Images?

- Ask a Radiologist
 - Answer: Medico-legal reasons
- Ask endoscopist why you **don't** keep your images
 - Answer: Medico-legal reasons!
- So we keep our images because there is a gold mine of additional data not just to document that the study was interpreted accurately!
- But how can we make all of that pixel data available for medical decision support as part of “Big Data”?

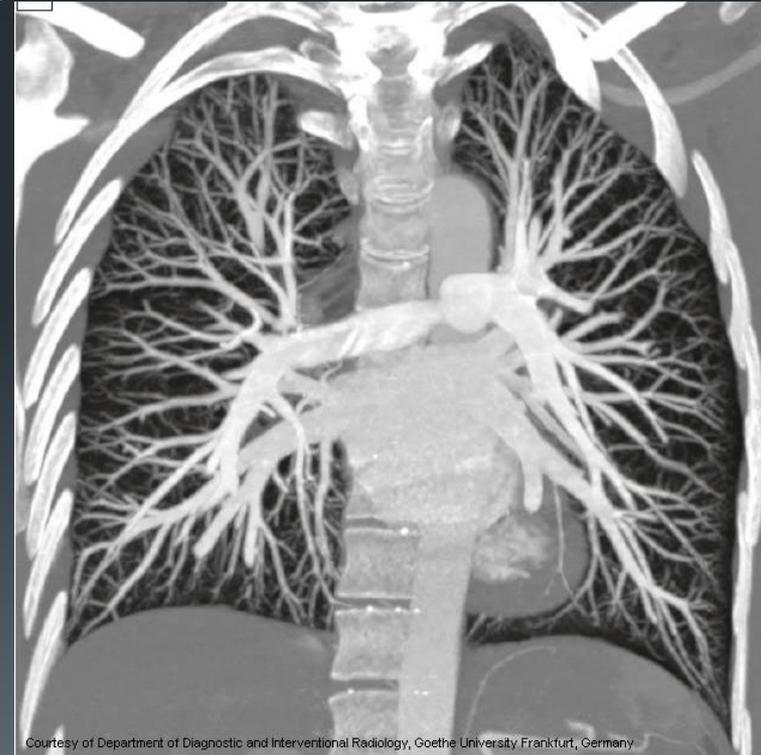
Imaging As “Physical Exam”- Capturing that “Dark Matter” Pipeline of Algorithms and then Store AIM/XML Tags...

- 1,000's of parameters can be discerned for future indexing and reference from a single, for example, CT pulmonary angiogram
 - Positive or negative for PE
 - Lung Nodules
 - Bone mineral density
 - Calcium score
 - Cardiac chamber size
 - **COPD Index**
 - Lung texture
 - Cystic changes
 - Pulmonary vascular assessment
 - Lung volumes
 - Liver size and texture
 - Pleural calcifications
 - Renal artery disease
 - Gallstones
 - 1000s more could be noted
 - by human or computer observer



Can Also Perform Automated Analysis of Image Quality in this “Pipeline”

- Automatic evaluation of
 - Timing of contrast bolus
 - Amount of contrast enhancement in selected vessels
 - Oral contrast quality in stomach and bowel
 - Image to noise ratio
 - Field of view
 - Motion artifacts
 - Metallic artifacts

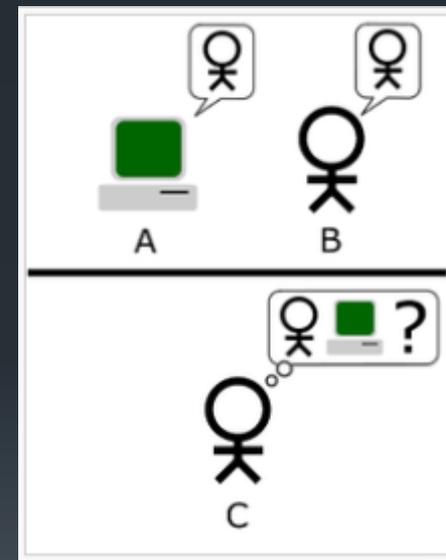


Capturing Data from Images is a Really Hard Task Maybe Currently the Hardest Task for Computers

Introduced by Alan Turing (Imitation Game) in his 1950 paper “Computing Machinery and Intelligence”

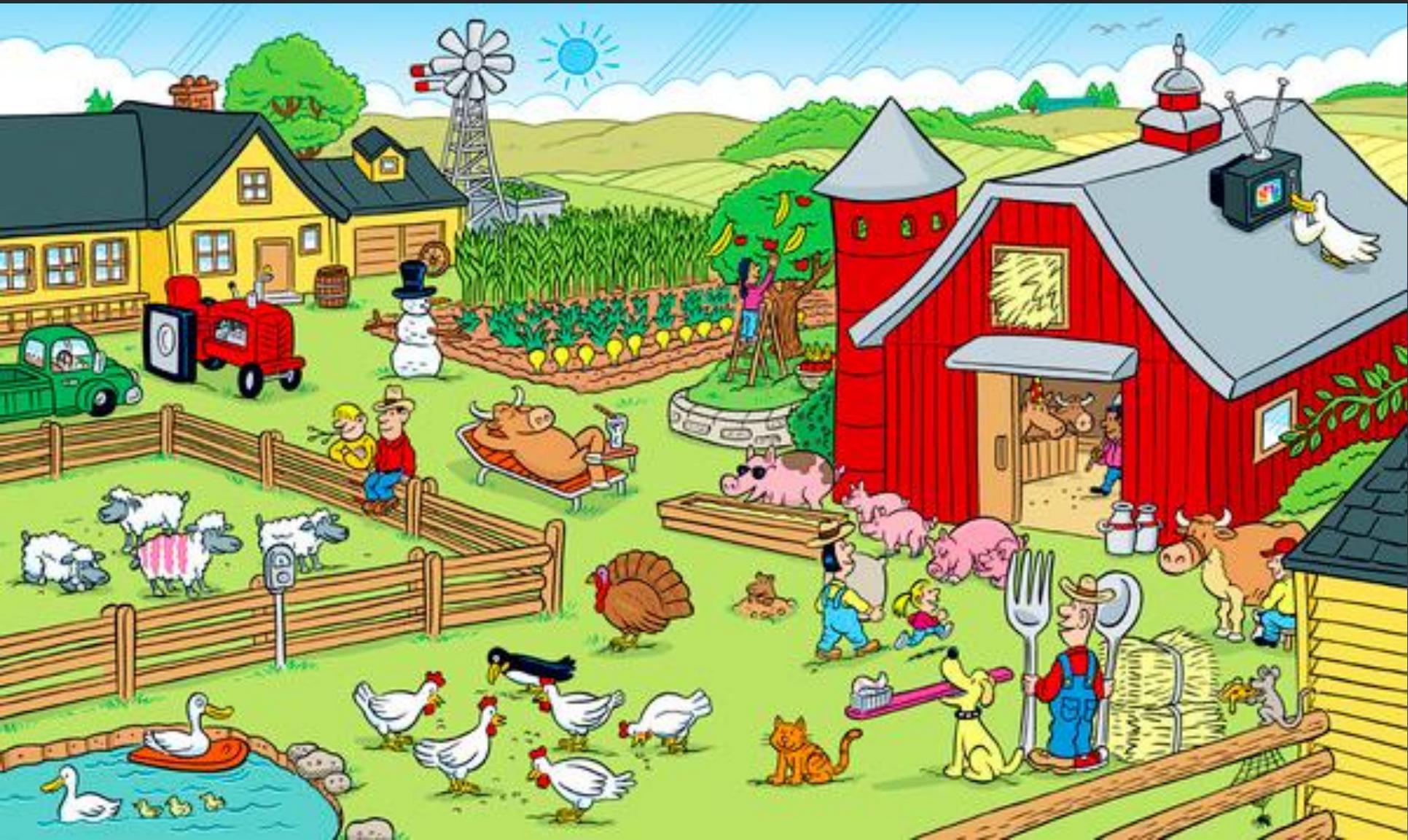
Opens with the words “I propose to consider the question, ‘Can machines think?’”

Asks whether a computer could fool a human being in another room into thinking it was a human being



Highlight's Magazine: What's Wrong with This Picture?

51



Ultimate Challenge: Medical Imaging

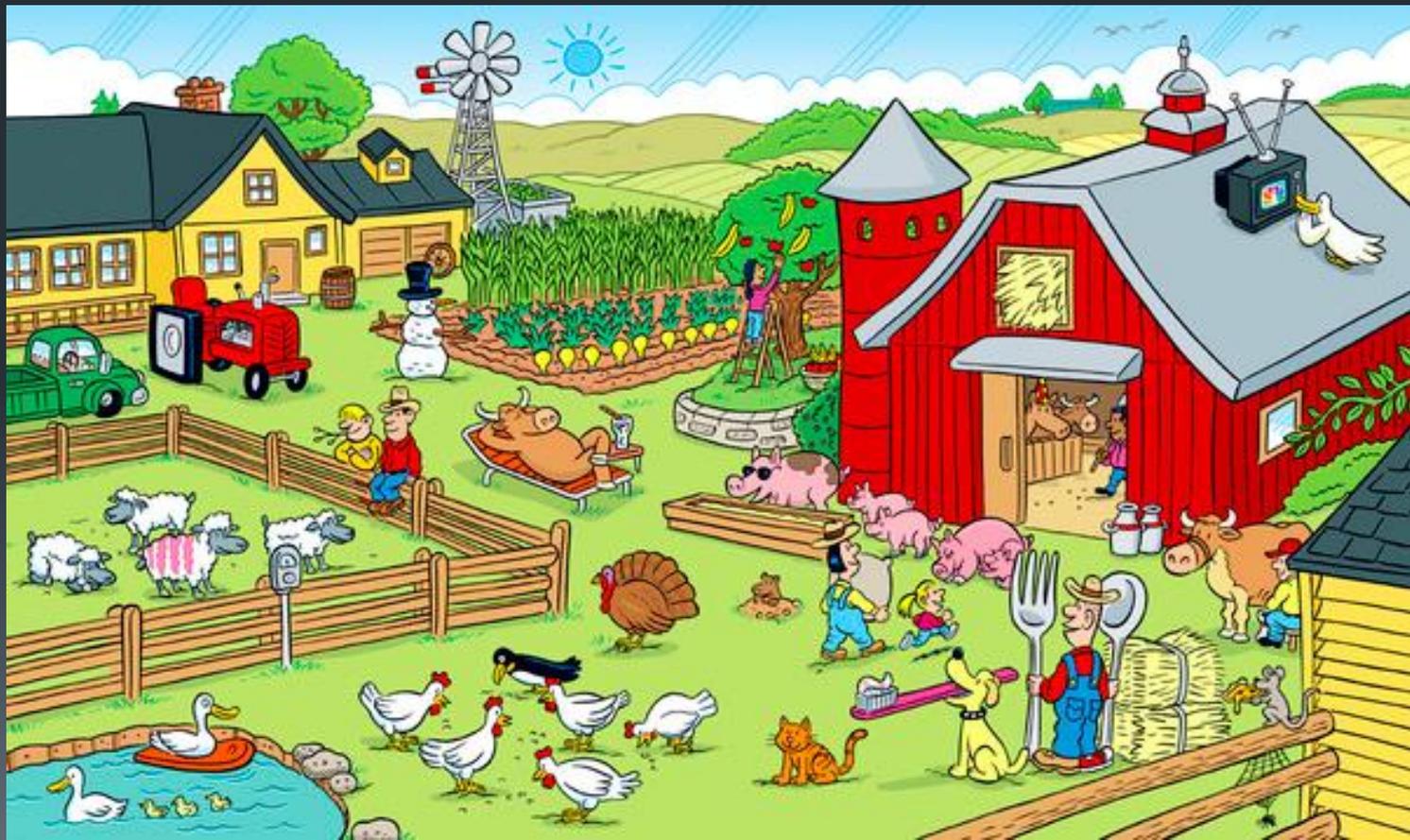
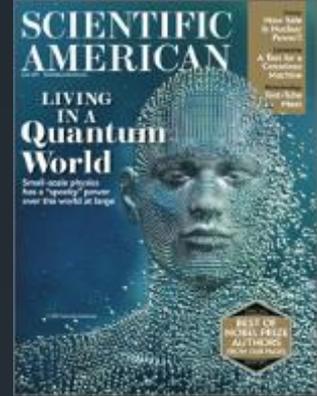
Scientific American June 2011

Testing for Consciousness

Alternative to Turing Test

Highlights for Kids “What’s Wrong with this Picture?”

Christof Koch and Giulio Tononi



Imaging May Be Ultimate/Future Frontier For “AI” Software



Using Big Data to Answer Big Questions and Little Questions in Diagnostic Imaging

- Examples of “little questions” requiring “little data” include those related to real time dashboards as well as scorecards
 - What is my report turnaround time?
 - Who are my most prolific referring clinicians?
 - How many unread studies are in my queue?
 - What is my patient waiting time
- Examples of “big questions” requiring “big data”
 - What is the impact of CT Pulmonary angiography on patient mortality and morbidity? Is it being over-utilized or underutilized?
 - Should CMS reimburse for CT screening exams in smokers over the age of 55?

Big Data Applications



- Auto-protocolling imaging studies personalized to a specific patient
- Automated diagnostic systems for brain tumors
- Personalized screening
- Personalized follow up recommendations (Personalized Fleischner guidelines for example)

Big Data Challenges

56

- Genomic tumor evaluation and correlation with imaging findings and prediction of most likely diagnosis
- Multi-parametric imaging analysis (e.g. MRI plus PET etc. for lymphoma etc.)
- Appropriateness criteria but based on empirical findings rather than expert recommendations
- Injection protocols optimized from all data and all previous injections
- Intelligent CAD applications

“Little Nodule, Little Question” Requiring Big Data: Clinical Case



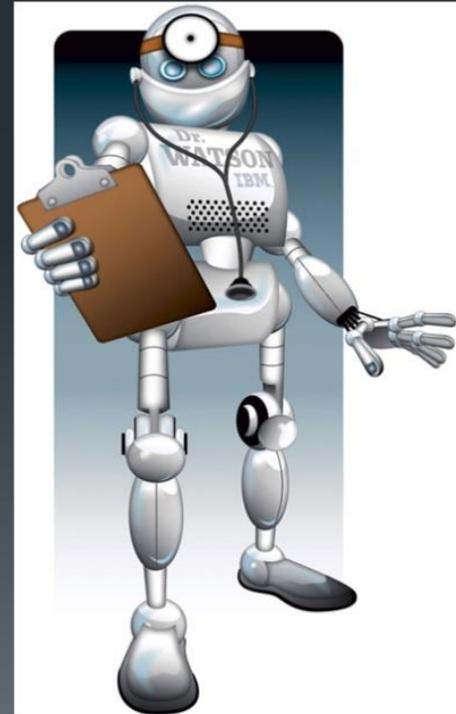
- Your next door neighbor and friend, Mr. Akami, a 62 year old native Hawaiian smoker with COPD who gets admitted for an elective Bunionectomy
- 6 mm spiculated soft tissue density right upper lobe nodule is discovered on “routine” pre-op exam and confirmed on CT with no other abnormalities
- What is the likelihood that it is malignant?
- How should this nodule be followed up?
- Can we use Big Data to solve these questions?



Year of Artificial Intelligence in Medicine



- 2011 will likely be remembered as the year of the re-emergence of artificial intelligence in medicine with Watson and of course, Siri, arguably the best feature of the new iPhone 4S and 5



Ask Siri “What are the chances that Mr. Akami Has Cancer”

Br J Radiol. Jul 2011; 84(1003): 661–668.
doi: [10.1259/bjr/24661484](https://doi.org/10.1259/bjr/24661484)

PMCID: PMC3473490

MRI in lung cancer: a pictorial essay

[B Hochegger](#), MD,¹ [E Marchiori](#), MD, PhD,² [O Sedlaczek](#), MD,³ [K Irion](#), MD, PhD,⁴ [C P Heussel](#), MD,² [S Ley](#), MD,² [J Ley-Zaporozhan](#), MD,² [A Soares Souza, Jr](#), MD, PhD,⁵ and [H-U Kauczor](#), MD²

[Author information](#) ► [Article notes](#) ► [Copyright and License information](#) ►

This article has been [cited by](#) other articles in PMC.

Abstract

Go to:

Imaging studies play a critical role in the diagnosis and staging of lung cancer. CT and 18-fluorodeoxyglucose positron emission tomography CT (PET/CT) are widely and routinely used for staging and assessment of treatment response. Many radiologists still use MRI only for the assessment of superior sulcus tumours, and in cases where invasion of the spinal cord canal is suspected. MRI can detect and stage lung cancer, and this method could be an excellent alternative to CT or PET/CT in the investigation of lung malignancies and other diseases. This pictorial essay discusses the use of MRI in

How About Watson's Deep Q/A Software for Mr. Akami?



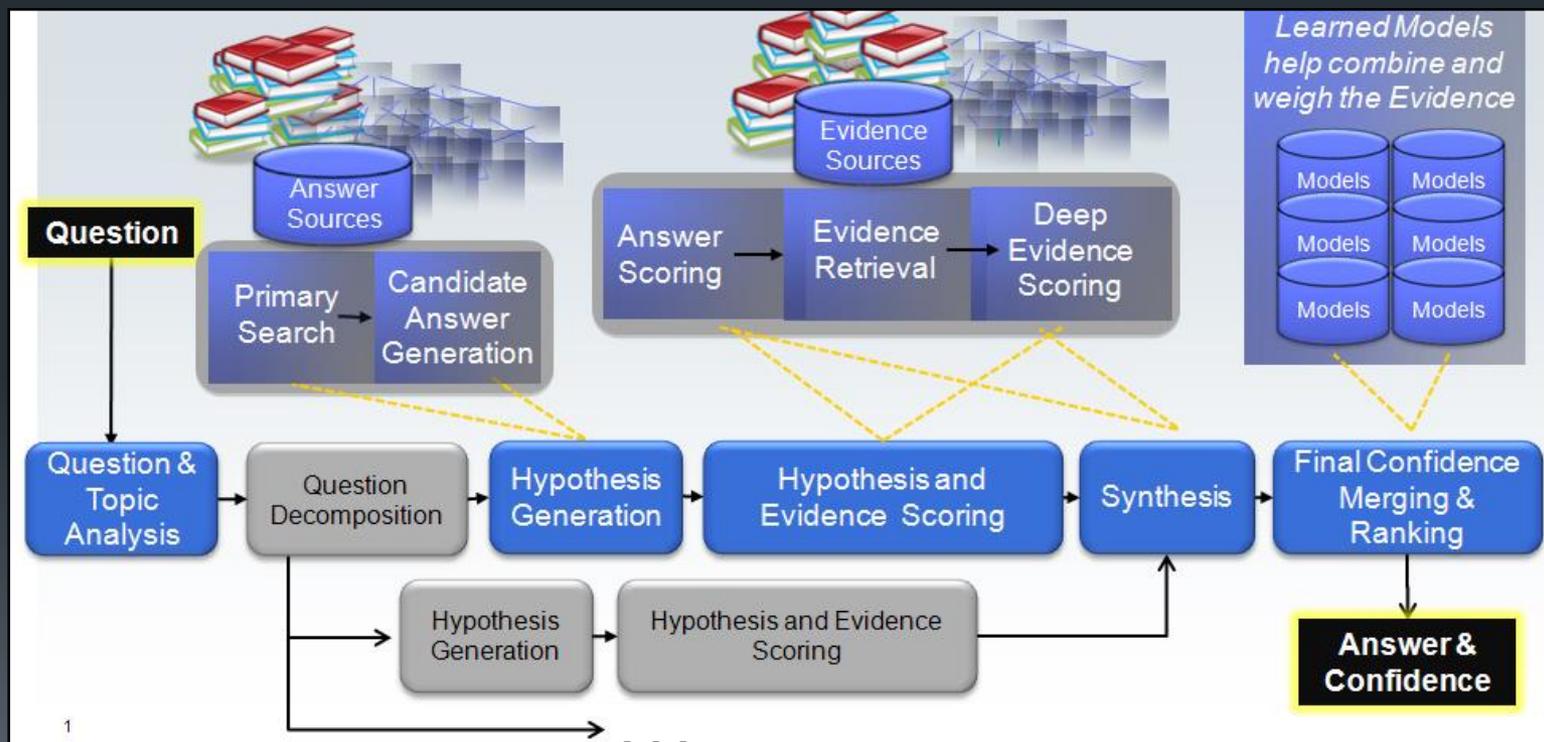
Watson is Fast

- Watson can process 500 gigabytes, the equivalent of a million books, per second
- Hardware cost has been estimated at about \$3 million
- 80 TeraFLOPs , 49th in the Top 50 Supercomputers list
- Content was stored in Watson's RAM for the game because data stored on hard drives too slow to process

Deep Q/A

- Does not map question to database of answers
- Represents software architecture for analyzing natural language content in both questions and knowledge sources
- Discovers and evaluates potential answers and gathers and scores evidence for those answers using unstructured sources such as natural language documents and structured sources such as relational and knowledge databases

High Level View of DeepQA Architecture





Challenges For Watson Deep Q/A

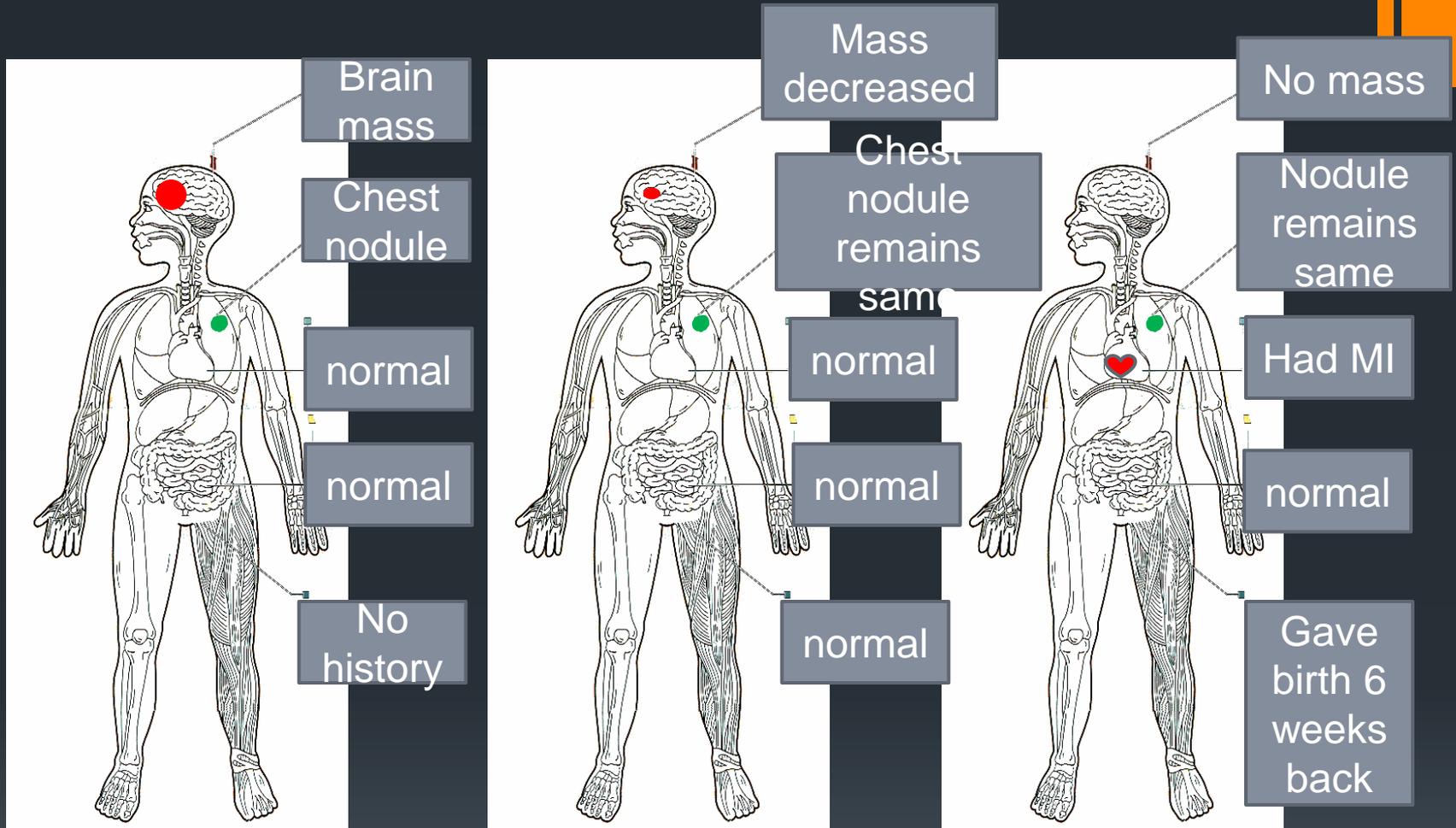
- Designed for questions and answers for Jeopardy!
- Coming up with most likely answer in Jeopardy! or even answer to medical board question is very different from actual diagnosis and treatment
 - Most patients have many diseases
 - There is no one correct gold standard answer
 - So word proximity, and recency, and other things that worked for Jeopardy are unlikely to work in Watson Medical Tool



Challenges for Watson Deep Q/A

- Does reasonably well at medical quiz questions such as internal medicine board examination
- Has difficulty with many aspects of the EMR such as abbreviations, cut and pasted text, templates, and so many of the things that make NLP such a huge challenge
- Biggest challenge may be lack of access to the gold mine of databases in radiology and pathology that have been collected over the many years such as NLST, DMIST, PLCO, etc.

Synthesis/Display of Complex Information in EMR



January 2000

May 2001

March 2002

Helping Mr. Akami

Can We Personalize Our Decision Making?

How Can We Make Radiology Images

Machine Intelligible?



Top Ten Informatics Challenges We Need to Address to Extract “Big Data” from Medical Images



TOP TEN List

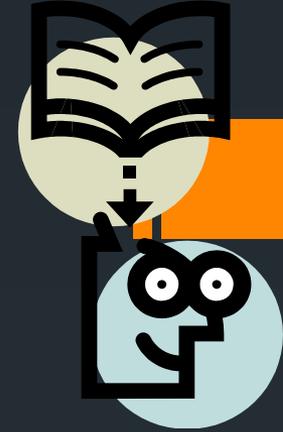
Page: 10 of 10

#10

Lack of Acquisition “Standards”

- Without an acquisition template or “standard” measurements will be of limited value
- We have demonstrated major differences in image analysis based on different acquisition parameters
- UPICT (uniform protocols in clinical trials)





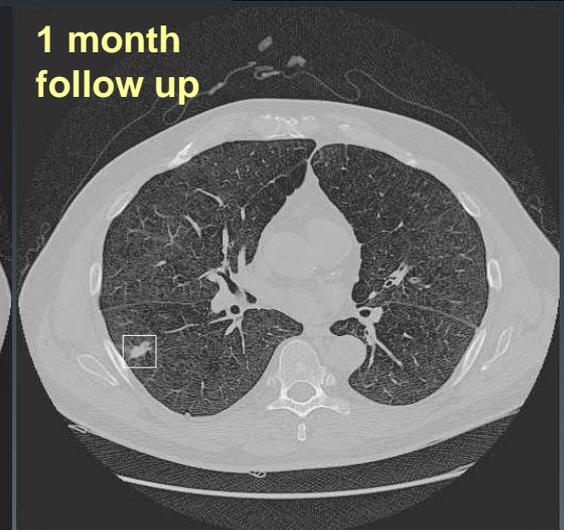
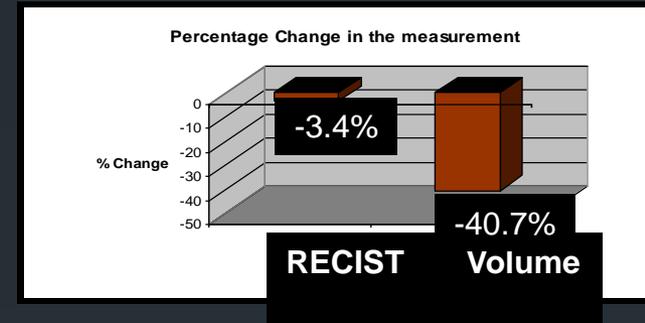
#9

Lack of a Radiology Lexicon

- Limited radiology terminology in Snomed CT (Systematized Nomenclature of Medicine Clinical Terms) or UMLS (Unified Medical Language System)
 - Current general medical lexicons only include about 20% of terms used in radiology reports
- Don't have consensus on acquisition parameters such as MRI sequences including GRASS, ROAST, etc. to describe acquisition standards

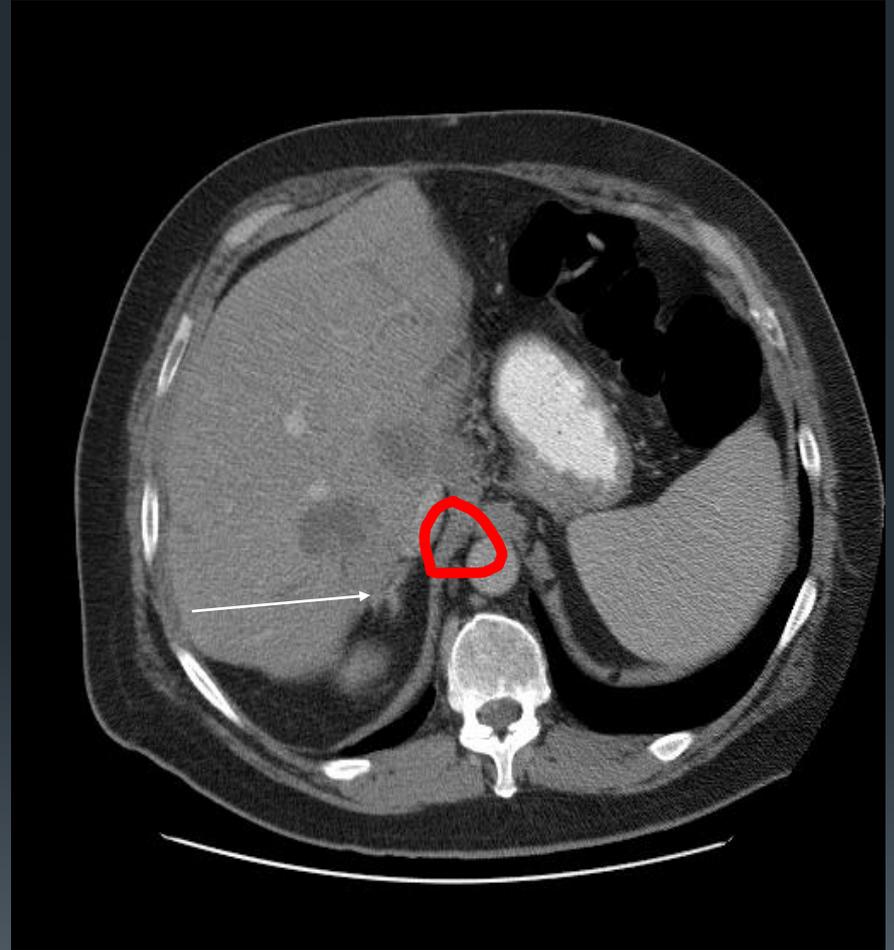
#8 Quantitative Imaging: Difficulty Measuring Lesion Size and Change over Time

- If imaging is to be successful as a biomarker it is critical that change in an image over time can be quantified
- Reference databases constructed to study this challenge are available on NCIA
 - LIDC (Lung Imaging Database Consortium)
 - Reference Image Database Resource (RIDER)



#7 Lack of Annotation and Mark-up Standards - Absolutely Critical to Success

- There is currently no consensus on image mark ups and annotation making anything done on one workstation unlikely to be accessible on another one
- This has largely been addressed by the AIM “standard” which uses standard vocabularies to tag portions of an image, whole images, or series or studies



AIM Annotation Workstation

Daniel Rubin MD - Stanford



The screenshot displays the AIM Annotation Workstation interface. On the left, a CT scan of a liver is shown with a green line indicating a lesion. A tooltip above the lesion reads: "lesion 1", "Length: 4.840 cm (84.281 pix)".

The main window, titled "IPAD", has a "Sheet" tab selected. It contains a table with the following data:

Status	ROI	Measurements	Findings	Locations
Valid	lesion1	Length: 9.4 cm	O: mass C: hypodense, irregularly sh... O: abscess C: almost certainly present	right lobe of liver left lobe of liver

Below the table, the text reads: "hypodense irregularly shaped mass; abscess almost certainly present in right lobe of liver, left lobe of liver".

At the bottom of the window, a status bar displays the following messages: "The knowledge base is loading...", "The annotation for lesion1 is valid and was saved.", and "The annotation for lesion1 is valid and was saved.".



#6 Difficulty in Acquiring and Submitting Cases in a Secure Way

- Clinical Picture Archiving and Communication Systems (PACS) are patient centric and are not designed to facilitate export of images
- This has resulted in the need for each group conducting clinical trials to “re-invent the wheel” with regard to software (typically proprietary) and often hardware solutions

#5 Multiple Image Interpretation Platforms with Different Software Makes Comparison of Results and Sharing of Results Difficult



extensible
IMAGING PLATFORM

- There is a wide variation in imaging software in quantitative analysis
- XIP (Extensible Imaging Platform)

#4

Lack of Standardized Reference Image Sets and Phantoms – NCI Archive Federated System

The screenshot shows the National Cancer Imaging Archive (NCIA) website in a browser window. The address bar displays <https://imaging.nci.nih.gov/ncia/faces/baseDef.tiles>. The page features a navigation menu with options like HOME, SEARCH IMAGES, MANAGE DATA BASKET, and HELP. A sidebar on the left contains quick links such as DICOM Image Viewers, NCIA NEWS, and NCIA USER'S GUIDE. The main content area includes a 'WELCOME TO NATIONAL CANCER IMAGING ARCHIVE' section with a description of the archive's purpose and a list of benefits. A 'USER LOGIN' section contains input fields for EMAIL and PASSWORD, along with Login and Register buttons. A 'NEW USER REGISTRATION' section provides instructions for new users. The footer includes links for CONTACT US, PRIVACY NOTICE, DISCLAIMER, ACCESSIBILITY, and SUPPORT, along with logos for the National Cancer Institute and FIRSTGov.

WELCOME TO NATIONAL CANCER IMAGING ARCHIVE

Welcome to the National Cancer Imaging Archive (NCIA). NCIA is a searchable repository of in vivo cancer images that provides the cancer research community, industry, and academia with access to image archives to be used in the development and validation of analytical software tools that support:

- Lesion detection and classification
- Accelerated diagnostic imaging decision
- Quantitative imaging assessment of drug response

NCIA provides access to imaging resources that will improve the use of imaging in today's cancer research and practice by:

- Increasing the efficiency and reproducibility of imaging cancer detection and diagnosis
- Leveraging imaging to provide an objective assessment of therapeutic response
- Ultimately enabling the development of imaging resources that will lead to improved clinical decision support.

USER LOGIN

EMAIL

PASSWORD

[Register](#)

NEW USER REGISTRATION

New users are asked to complete a one-time registration form. Please make note of your username and password for future visits. Registration will ensure that we can inform you about important changes and new data.

#3 Disconnect Between XML Restful Interfaces and DICOM and HL7

- Middleware project caBIG in vivo Imaging Workspace

#2 Patient-centric Electronic Medical Record and PACS makes Data Mining Difficult

- Need to work with designers of the patient electronic medical record and PACS to begin to think of ways to index and search through data without compromising patient privacy and security

#1

Challenge to Promote Sharing of Images and Related Data

- Images associated with clinical trials should be available for users after the study is completed and when approved by the principal investigators, while the study is underway
- This could make images available to investigators and industry without the huge expenses and amount of time and dollars typically required for a study involving clinical images

NEJM “Open Data” Drazen IOM Committee

The NEW ENGLAND JOURNAL of MEDICINE

EDITORIAL



Open Data

Jeffrey M. Drazen, M.D.

In the fall of 2013, the Institute of Medicine (IOM) convened a committee, on which I serve, to examine the sharing of data in the setting of clinical trials. The committee is charged with reviewing current practices on data sharing in the context of randomized, controlled trials and with making recommendations for future data-sharing standards. Over the past few months, the committee has prepared a draft report that reviews current practices on data sharing and lays out a number of potential data sharing models. Full

Open-data advocates argue that all the study data should be available to anyone at the time the first report is published or even earlier. Others argue that to maintain an incentive for researchers to pursue clinical investigations and to give those who gathered the data a chance to prepare and publish further reports, there should be a period of some specified length during which the data gatherers would have exclusive access to the information. Since these researchers could always agree to collaborate with oth

IOM Request for Comments



INSTITUTE OF MEDICINE

OF THE NATIONAL ACADEMIES

Board on Health Sciences Policy

[Committee on Strategies for Responsible Sharing of Clinical Trial Data](#)

STATEMENT OF TASK

An ad hoc committee of the Institute of Medicine will conduct a study to develop guiding principles and a framework (activities and strategies) for the responsible sharing of clinical trial data. For the purposes of the study, the scope will be limited to interventional clinical trials and “data sharing” will include the responsible entity (data generator) making the data available via open or restricted access*, or exchanged among parties. For the purposes of this study, data generator will include industry sponsors, data repositories, and researchers conducting clinical trials.

Specifically, the committee will:

- Articulate guiding principles that underpin the responsible sharing of clinical trial data.
- Describe a selected set of data and data sharing activities, including, but not limited to:
 - Types of data (e.g., summary, participant)
 - Provider(s) and recipient(s) of shared data
 - Whether and when data are disclosed publicly, with or without restrictions, or exchanged

- 
- Francis Collins, acknowledging the current “unhealthy” state of affairs with the whole NIH grant and review process and of course the value of sharing and reusing data
 - I’ve been a big advocate of the **Data Discovery Index** and we should have some discussion and make some suggestions about how this could apply to what types of information about an imaging data set would be important to index
 - Also, being able to access images using an API /REST/DICOM interface would be really valuable rather than just accessing these via a custom and idiosyncratic portal

Radiomics: Involves Tagging Images To Support Phenotype/Genotype Clinical Analysis

Image



Image Annotation Data

Imaging Observations: Definition of the Enhancing Margin Well-defined

Imaging Observations: Cysts No

Imaging Observations: Enhancement Quality: Mark/Avid

Imaging Observations: Calvarial Remodeling No

Imaging Observations: Cortical Involvement Yes

Imaging Observations: Thickness of the Enhancing Margin Thick/normal

Imaging Observations: Enhancing Tumor Crosses Midline No

Or and

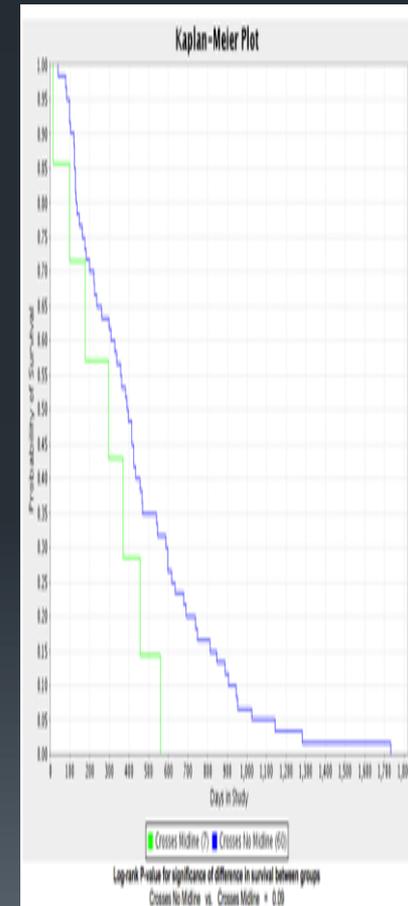
Genomic & Clinical

caIntegrator

Search TOCA Radiology caIntegrator

SubjectID	SampleID	Gene	Copy	Loss	Gain	Copy	Loss	Gain	Loss	Gain	Loss	Gain
TC2A0198	TC2A0198A	708	-0.0	1.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0
TC2A0192	TC2A0192A	157	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
TC2A0193	TC2A0193A	418	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
TC2A0195	TC2A0195A	110	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
TC2A0190	TC2A0190A	110	-0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
TC2A0199	TC2A0199A	128	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
TC2A0197	TC2A0197A	110	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
TC2A0194	TC2A0194A	110	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
TC2A0196	TC2A0196A	110	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
TC2A0195	TC2A0195A	110	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
TC2A0193	TC2A0193A	110	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Analytic Results



Template Guides Radiologist Through Findings in Structured Way and then Saves to AIM Data Service And Search is Made for Similar Patients

The screenshot displays the caIntegrator web interface for a brain MRI study. The browser address bar shows the URL: <https://caintegrator2.nci.nih.gov/caintegrator2/manageQuery.action>. The interface includes a list of structured observations on the left, a central MRI image, and a list of actions on the right.

Structured Observations:

- Imaging Observations: Definition of the Enhancing Margin Well-defin
- Imaging Observations: Cysts No
- Imaging Observations: Enhancement Quality : Mark/Avid
- Imaging Observations: Calvarial Remodeling No
- Imaging Observations: Cortical involvement Yes
- Imaging Observations: Thickness of the Enhancing Margin Thick/noc
- Imaging Observations: Enhancing Tumor Crosses Midline No

Actions:

- Remove

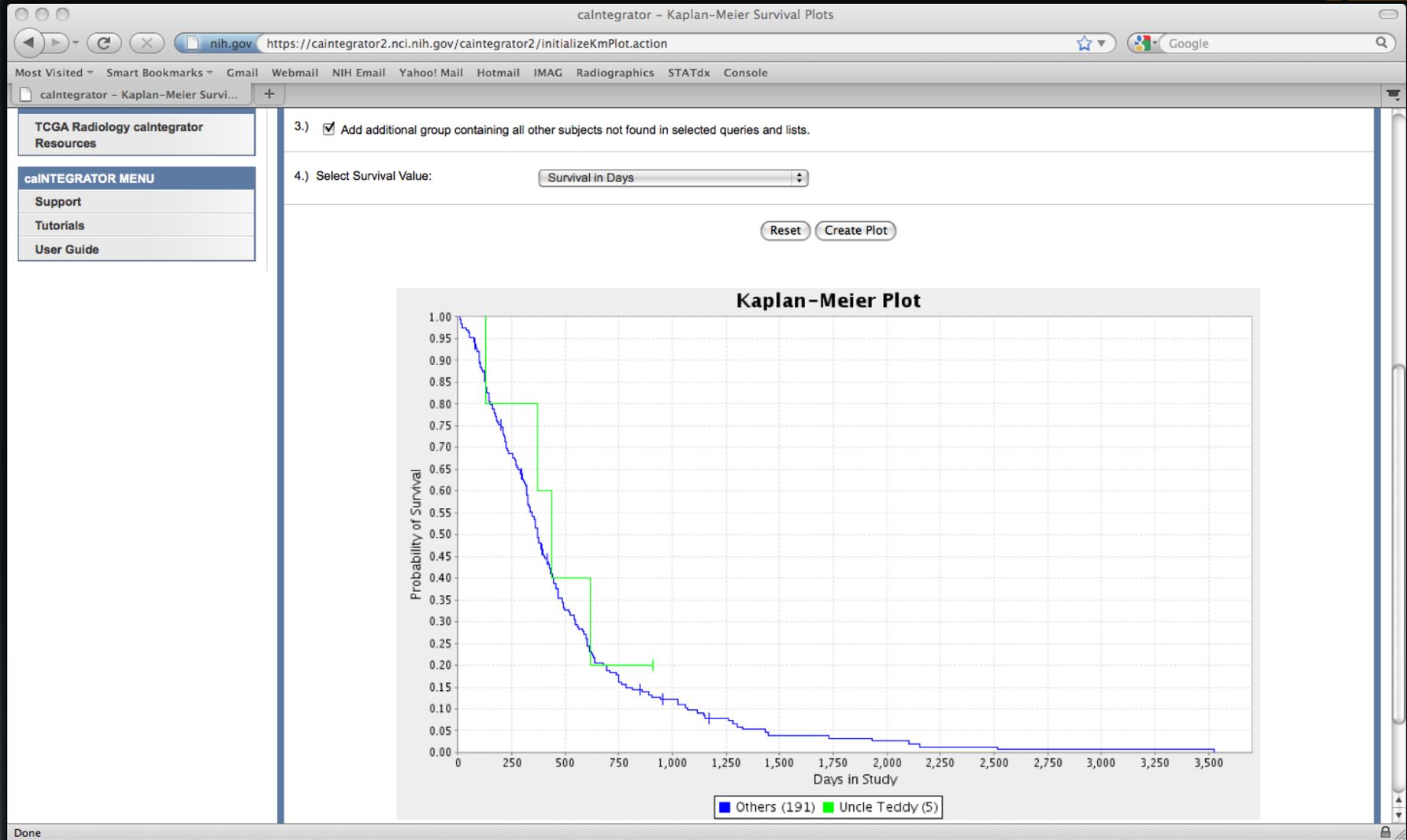
Navigation:

or and

The central image is an axial MRI scan of the brain showing a central enhancing lesion. Red arrows point from the following observations to the corresponding features on the MRI:

- Definition of the Enhancing Margin Well-defin
- Enhancement Quality : Mark/Avid
- Calvarial Remodeling No
- Cortical involvement Yes
- Thickness of the Enhancing Margin Thick/noc
- Enhancing Tumor Crosses Midline No

KM Plot For Specific Patient Based on TCGA Database



*Challenge is that Radiology Data is
High Dimensional*



VASARI MRI Parameters Just Scratching the Surface of Diagnostic Imaging Parameters



- MRI

- T1, T2, Proton density, Echo planar, Inversion Recovery, Perfusion, Diffusion-Kurtosis, Spectroscopy, many contrast agents and many sequences

- US

- Contrast, elastography, flow

- CT

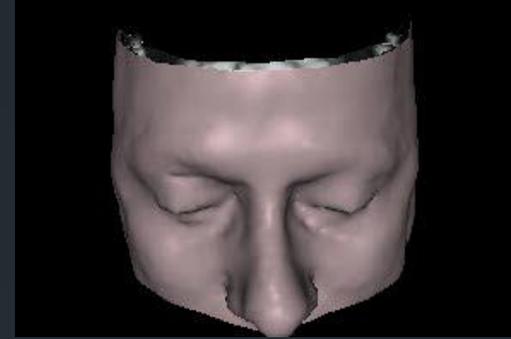
- Dynamic contrast
- Multi-spectral



PET Parameters

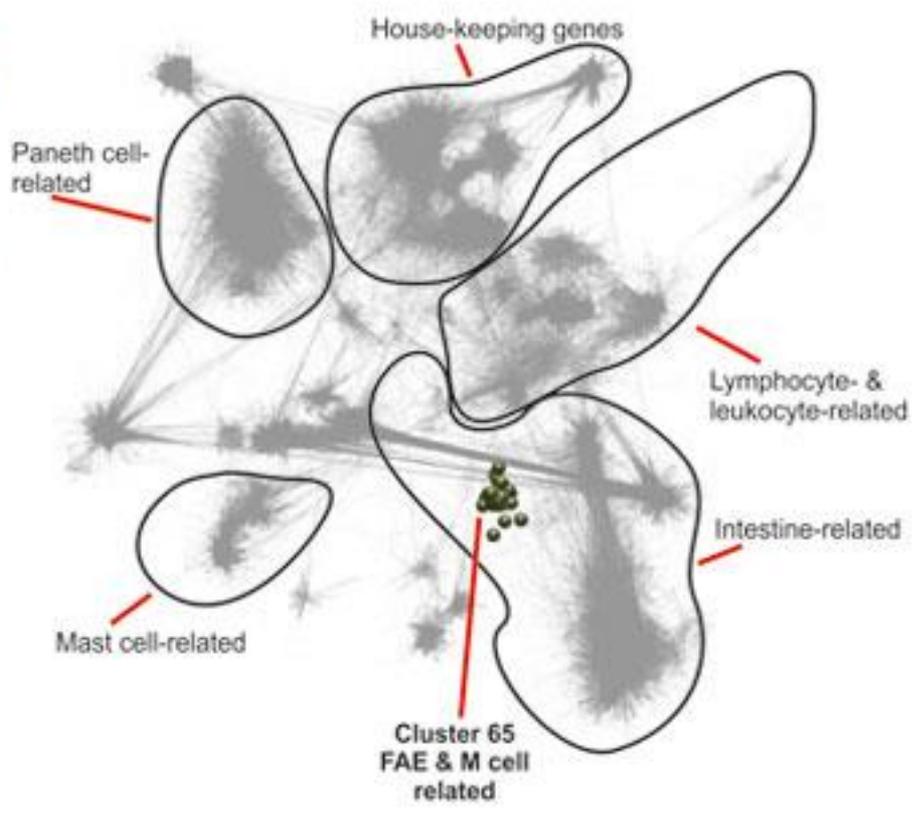
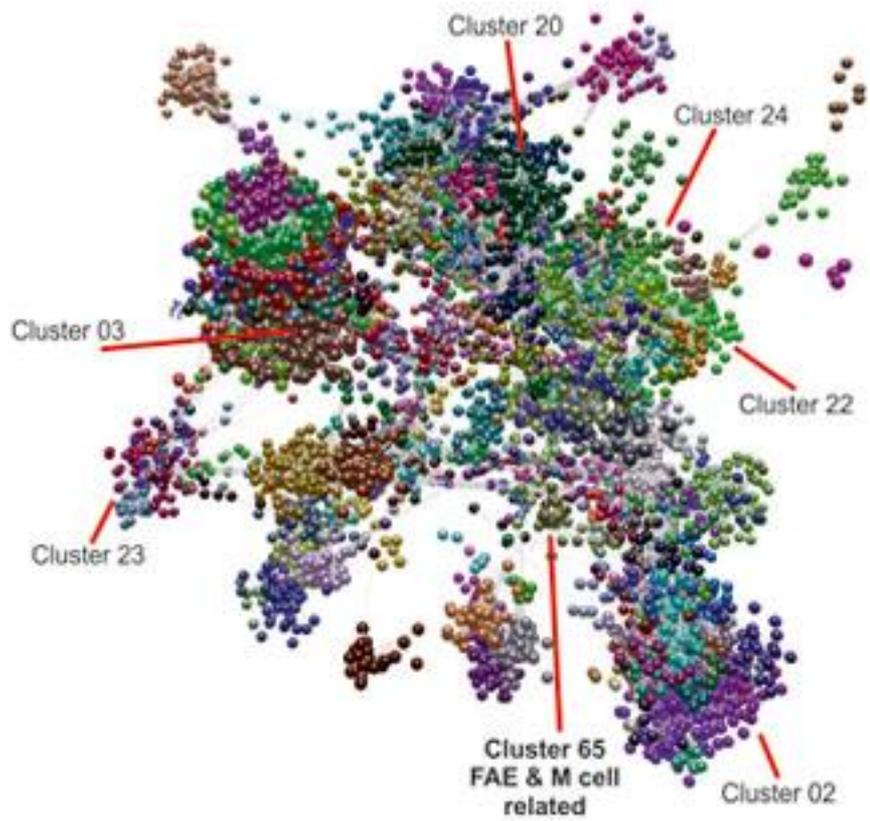
- Glucose utilization: FDG,
- Tumor Cellular Proliferation: ^{18}F -FLT-PET and tumor cellular proliferation
- Tumor Hypoxia: [^{18}F]fluoromisonidazole (F-MISO)
- Apoptosis: [^{18}F]ICMT-11
- [^{18}F]fluoroethyl-l-tyrosine (FET), [^{18}F]fluoro- α -methyltyrosine (FMT),, 6-[^{18}F]fluoro-dihydroxy-l-phenylalanine (F-DOPA), [^{11}C]choline (CHO) and [^{18}F]choline.





CT Exam as Physical Exam Almost Infinite Number of Parameters Including Contrast, Temporal, Functional

- 3D Image of Chest or heart or abdomen
- Single CT image 15,000 images then compare to previous studies and plot motion differences point to point and compare with clinical and lab and genomic/proteomic, etc.
- DICOM just subset of original sinogram data



High-Dimensional Informatics/Statistics are Much More than Just Crunching More Numbers with a Faster Computer and Bigger Spreadsheet

- Informatics of datasets having a greater number of dimensions than classically considered in traditional multivariate analysis
 - Dimension of the data vectors may be even larger than the sample size!





Need New Approaches to Analysis of High Dimensional Datasets

- In traditional statistical data analysis, we think of observations of instances of particular phenomena
- In traditional statistical methodology, such as regression models, we assumed many observations and a few, well chosen variables
- The trend today is towards more observations but even more so, to radically larger numbers of variables

- 
- Classical statistical methods are simply not designed to cope with this kind of explosive growth of dimensionality of the observation vector
 - In the coming century, high-dimensional data analysis will be a very significant activity, and completely new methods of high-dimensional data analysis will be developed; we just don't know what they are yet

High Dimensional Data Statistics \neq Data Mining



- Regression modeling – limited in high dimensional data analysis
 - Linear: Multiple variables are used to predict a “quantitative response” variable, uses linear algebra
 - Non-Linear: involves local linear fits, neural networks, radial basis functions, etc.

What Are High-Dimensional Informatics/Statistics?



- Informatics of datasets having a greater number of dimensions than classically considered in traditional multivariate analysis
 - Dimension of the data vectors may be even larger than the sample size!



The Curse of High Dimensionality

- Apparent intractability of understanding and visualizing and accurately approximating a general high-dimensional function
- The apparent intractability of integrating a high-dimensional function.



Blessings of Dimensionality

- Include the “concentration of measure phenomenon”
 - Means that certain random fluctuations are very well controlled in high dimensions

Big Data Analytic Techniques Non-⁹⁹ Statistics Based

■ Latent Variables Analysis

- Looks to uncover “latent variables” as responsible for a pattern that is seen in an array to discover important insights
- **Principal Component Analysis (PCA)** is an example which is widely used
 - Has been used in image analysis, e.g. facial recognition
 - Latent semantic indexing: uses PCA analysis to perform Web searches
 - Independent Component Analysis (ICA) developed in past several years on challenges such as EKG or EEG analysis
- **Cluster Analysis**
 - Used in analysis of genomic data

Analysis of High Dimensional Big Data Such as EMR Likely Requires Similar Tools and Approaches as are used In Medical Image Analysis

100

- No medical experts have really stepped up to the plate yet to help solve the really big, high dimensional data mining challenges in the EMR to look for new patterns and do real time personalized medicine for real time patient care, e.g. what statin should my patient be on given his lab profile, family history, history of diabetes, previous response to specific statins and the entire database of responses to statins?

Analysis of High Dimensional Big Data Such as EMR Likely Requires Similar Tools and Approaches as are used In Medical Image Analysis

- The Medical Imaging community routinely use tools for feature extraction, segmentation, and analysis that could prove invaluable for on the fly cohort selection, cohort evaluation and analysis required for routine medical care such as choice of optimal medications (statins, antibiotics, hypertensive meds, etc.)

Major Challenge for Next Generation Decision Support: Lack of Access to Large Databases

102



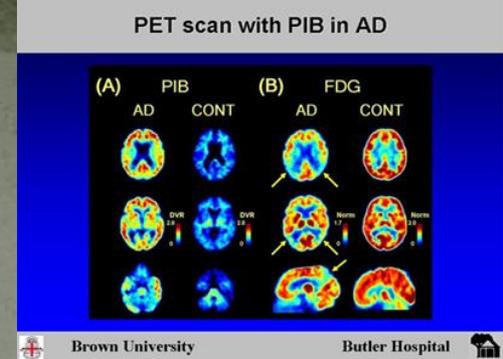
Discovering Other Untapped, Disconnected Gold Mines of Clinical and Research Data

- Despite all of the advances in computer technology we are arguably still at the paper stage of research as far as ability to discover and combine important data
 - Research data including those associated with major medical journals and clinical trials are typically created for a single purpose and beyond a one or two manuscripts, remain largely locked up or inaccessible
 - Even when the data are made accessible, they are typically associated with limited access through a proprietary Internet portal or even by requesting data on a hard drive
 - Often requires submission of a research plan and data and then a considerable wait for permission to use the data which is often not granted

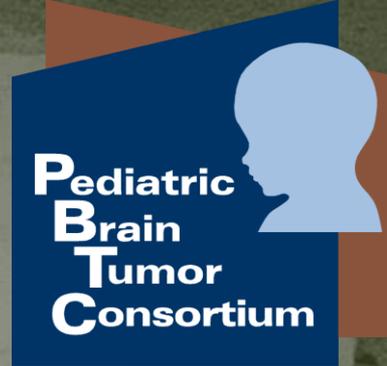


ADNI

- Alzheimer's Disease Neuroimaging Initiative
- Excellent example of patient data and associated images with great sharing model
- However requires access through their own portal and requires permission from ADNI Data Sharing and Publications Committee



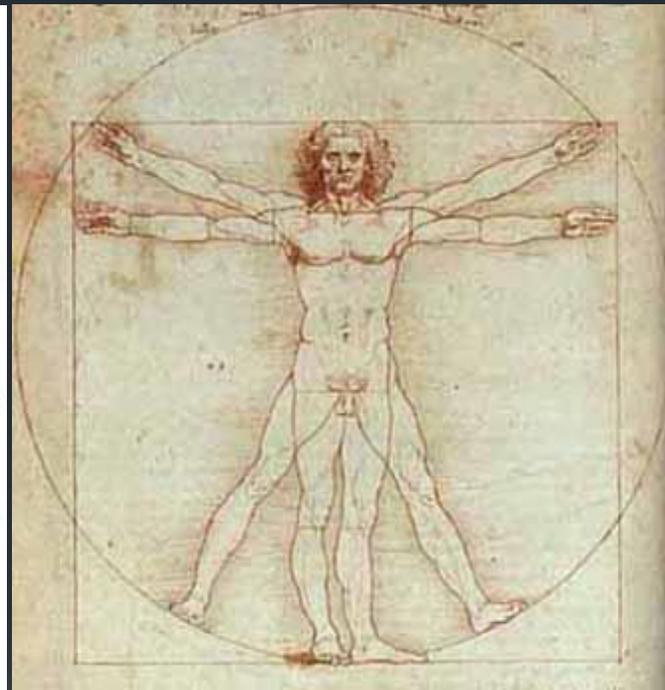
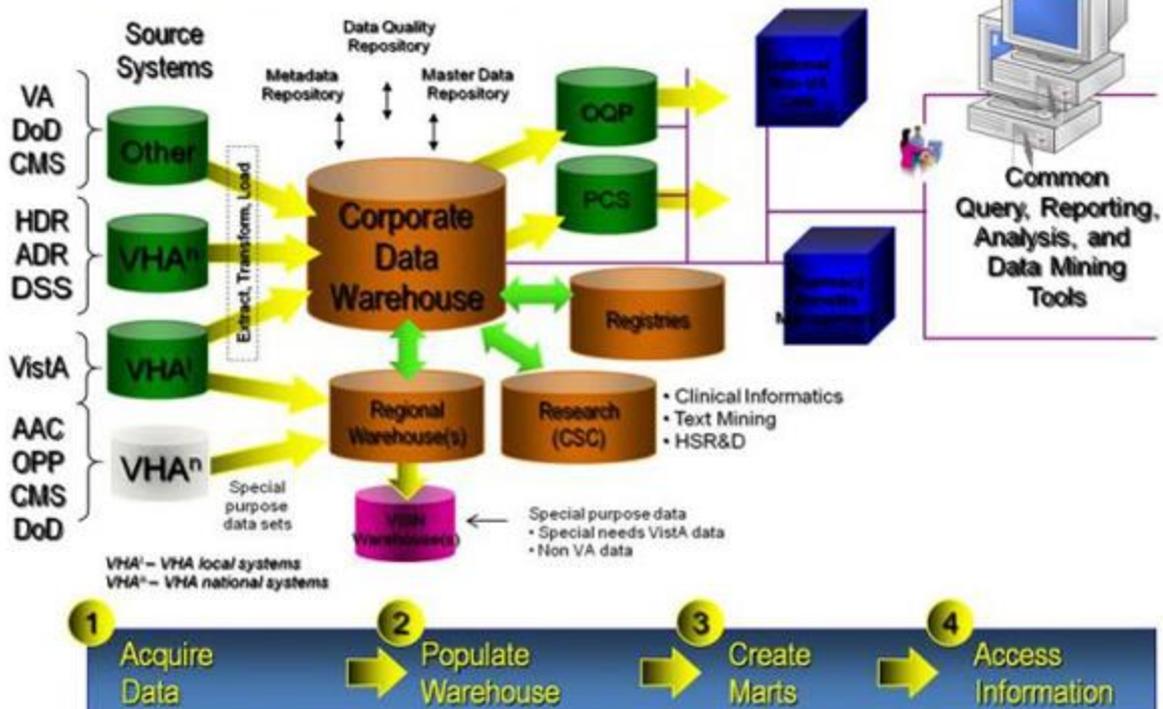
CTEP (NIH Cancer Therapy Evaluation Program) Pediatric Brain Tumor Consortium One of the Better Sources of Data



- As an NCI funded Consortium, the Pediatric Brain Tumor Consortium (PBTC) is **required** to make research data available to other investigators for use in research projects
- An investigator who wishes to use individual patient data from one or more of the Consortium's completed and published studies must submit in writing:
 - Description of the research project
 - Specific data requested
 - List of investigators involved with the project
 - Affiliated research institutions
 - Copy of the requesting investigator's CV must also be provided.
- The submitted research proposal and CV shall be distributed to the PBTC Steering Committee for review
- Once approved, the responsible investigator will be required to complete a Material and Data Transfer Agreement as part of the conditions for data release
- Requests for data will only be considered once the primary study analyses have been published

Institutional Database: VA's Corporate Data Warehouse Vinci

VHA Data Warehousing Visual Architecture – Current Vision



Discovering and Consuming Databases

- At best, freely sharable databases are accessed using their own idiosyncratic web portal
- Currently no index of databases or their content
- No standards exist to describe how databases can “advertise” their content and availability (free or business model) and their data provenance and sources and peer review, etc.
- Would be wonderful project for NLM to investigate the creation of an XML standard for describing the content of databases
- This will be critical to the continuing success of the Watson project in my opinion

Back to Mr. Akami

- Your next door neighbor and friend, Mr. Akami, a 62 year old native Hawaiian smoker with COPD who gets admitted for an elective Bunionectomy
- 6 mm spiculated soft tissue density right upper lobe nodule is discovered on “routine” pre-op exam and confirmed on CT with no other abnormalities
- What is the likelihood that it is malignant?
- How should this nodule be followed up?
- **First question is quantitative one whether the nodule is actually 6 mm or not**



Recent CMS Approval for Reimbursement of Screening for Lung Cancer in Smokers

- CMS memo from early February 2015 approves what will amount to about \$1.9 billion per year for reimbursement of annual low-dose lung cancer screening for asymptomatic individuals ages 55 to 77 years, who have a tobacco smoking history of at least 30 pack-years
- Screening participants must be a current smoker or have quit within the past 15 years

Fleischner Society Guidelines

Recommendations for Follow-up and Management of Nodules Smaller than 8 mm Detected Incidentally at Nonscreening CT

Nodule Size (mm)*	Low-Risk Patient†	High-Risk Patient‡
≤4	No follow-up needed§	Follow-up CT at 12 mo; if unchanged, no further follow-up
>4–6	Follow-up CT at 12 mo; if unchanged, no further follow-up	Initial follow-up CT at 6–12 mo then at 18–24 mo if no change
>6–8	Initial follow-up CT at 6–12 mo then at 18–24 mo if no change	Initial follow-up CT at 3–6 mo then at 9–12 and 24 mo if no change
>8	Follow-up CT at around 3, 9, and 24 mo, dynamic contrast-enhanced CT, PET, and/or biopsy	Same as for low-risk patient

Note.—Newly detected indeterminate nodule in persons 35 years of age or older.

* Average of length and width.

† Minimal or absent history of smoking and of other known risk factors.

‡ History of smoking or of other known risk factors.

§ The risk of malignancy in this category (<1%) is substantially less than that in a baseline CT scan of an asymptomatic smoker.

|| Nonsolid (ground-glass) or partly solid nodules may require longer follow-up to exclude indolent adenocarcinoma.

National Lung Screening Trial Dataset and Decision Support Project

- Can we personalize the Fleischner criteria using data from the National Lung Screening Trial?
- Could the criteria for follow-up be refined and personalized more than high risk smoker vs. lower risk patient based on:
 - Geographic location?
 - Patient age/sex?
 - Characteristics of nodule e.g. shape (spiculated or smoothly rounded), containing calcification?
 - Presence of additional nodules?

NLST Personalized Lung Nodule Diagnosis and Treatment

NLST Nodule Search

Patient Characteristics:

Age (54-75): 61 - 63

Gender:

- Male
- Female
- All

Smoking Years:

10 - 68

Smoking Pack Years:

15 - 567

Nodule Characteristics:

Size: 7 - 9 mm

Margins:

- Spiculated (Stellate)
- Smooth
- Poorly defined
- Undetermined

Opacity:

- Soft tissue
- Ground glass
- Mixed
- Fluid/water
- Fat
- Other
- Undetermined

Lobe Location:

All | None

- Right Upper
- Right Middle
- Right Lower
- Left Upper
- Lingula
- Left Lower
- Other (Crosses Boundaries)

Find Nodules

Query Results:

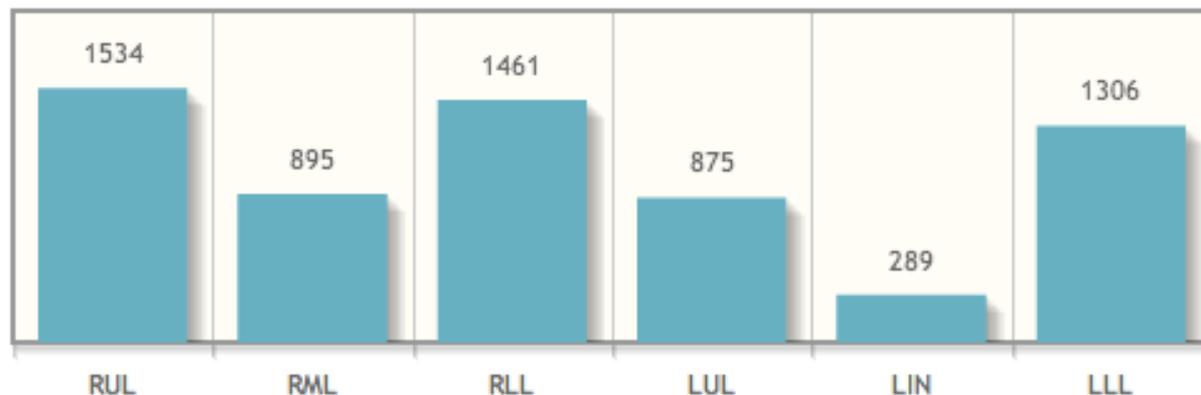
For the patient and nodule criteria selected:

Patient/Nodule Matches: 72

Number of cancers in this subset: 11 (15% of nodules)

Mean follow up till cancer diagnosis: 3 years +/- 1.5 years
(Range: 0 to 5 years)

Total nodules (Age: 61 - 63): 6450



Nodules by Lobe

5% of Nodules for Males 60 to 65 that were 5-7mm were Malignant

Patient Characteristics:

Age: 60 - 65



Gender:

- Male
 Female
 All

Smoking Years:

10 - 68



Smoking Pack Years:

15 - 567



Nodule Characteristics:

Size: 5 - 7 mm



- Include
 Micronodules

Margins:

All | None

- Spiculated (Stellate)
 Smooth
 Poorly defined
 Undetermined

Opacity:

All | None

- Soft tissue
 Ground glass
 Mixed
 Fluid/water
 Fat
 Other
 Undetermined

Lobe Location:

All | None

- Right Upper
 Right Middle
 Right Lower
 Left Upper
 Lingula
 Left Lower
 Other (Crosses
 Boundaries)

Screening Year

All | None

- Year 1
 Year 2
 Year 3

Find Nodules

Results returned in 1.153 seconds

Stats

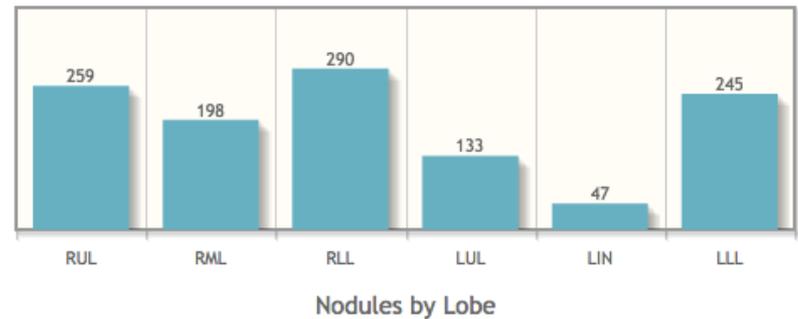
Patient/Nodule Matches: 1180

Number of cancers in this subset: 60 (5% of nodules)

Mean follow up till cancer diagnosis: 1 years \pm 1.6 years (Range: 0 to 6 years)

Distribution

Total Nodules (Age: 60 - 65): 1180



Fleischner Criteria

Category	Nodules	Fleischner Recs
≤ 4 mm	0, 0 total cancers (NaN%)	Followup CT at 12 months; if unchanged, no further follow-up

However if **Spiculated** rather than smooth then **11%** of Nodules for Males 60 to 65 that were 5-7mm were Malignant (mostly RUL)

Patient Characteristics:

Age: 60 - 65



Gender:

- Male
- Female
- All

Smoking Years:

10 - 68



Smoking Pack Years:

15 - 567



Nodule Characteristics:

Size: 5 - 7 mm



- Include Micronodules

Margins:

All | None

- Spiculated (Stellate)
- Smooth
- Poorly defined
- Undetermined

Opacity:

All | None

- Soft tissue
- Ground glass
- Mixed
- Fluid/water
- Fat
- Other
- Undetermined

Lobe Location:

All | None

- Right Upper
- Right Middle
- Right Lower
- Left Upper
- Lingula
- Left Lower
- Other (Crosses Boundaries)

Screening Year

All | None

- Year 1
- Year 2
- Year 3

Find Nodules

Results returned in 1.029 seconds

Stats

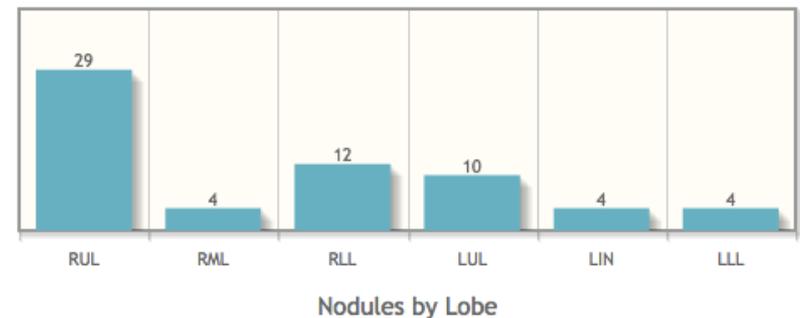
Patient/Nodule Matches: 63

Number of cancers in this subset: 7 (11% of nodules)

Mean follow up till cancer diagnosis: 2 years ± 1.4 years (Range: 0 to 4 years)

Distribution

Total Nodules (Age: 60 - 65): 63



Fleischner Criteria

Category	Nodules	Fleischner Recs
<= 4 mm	0, 0 total cancers (NaN%)	Followup CT at 12 months; if unchanged, no further follow-up

Of those in the RUL, 13% of Nodules for Males 60 to 65 that were 5-7mm were Malignant

Patient Characteristics:

Age: 60 - 65



Gender:

- Male
- Female
- All

Smoking Years:

10 - 68



Smoking Pack Years:

15 - 567



Nodule Characteristics:

Size: 5 - 7 mm



- Include Micronodules

Margins:

All | None

- Spiculated (Stellate)
- Smooth
- Poorly defined
- Undetermined

Opacity:

All | None

- Soft tissue
- Ground glass
- Mixed
- Fluid/water
- Fat
- Other
- Undetermined

Lobe Location:

All | None

- Right Upper
- Right Middle
- Right Lower
- Left Upper
- Lingula
- Left Lower
- Other (Crosses Boundaries)

Screening Year

All | None

- Year 1
- Year 2
- Year 3

Find Nodules

Results returned in 0.97 seconds

Stats

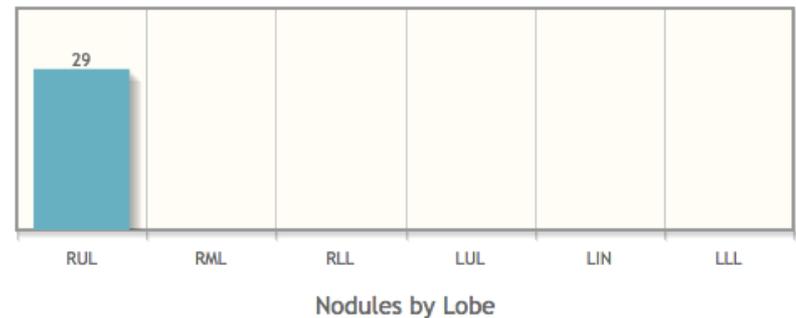
Patient/Nodule Matches: 29

Number of cancers in this subset: 4 (13% of nodules)

Mean follow up till cancer diagnosis: 2 years \pm 1.2 years (Range: 0 to 3 years)

Distribution

Total Nodules (Age: 60 - 65): 29



Fleischner Criteria

Category	Nodules	Fleischner Recs
≤ 4 mm	0, 0 total cancers (NaN%)	Followup CT at 12 months; if unchanged, no further follow-up

Opportunities for Screening In Diagnostic Imaging

- CMS recently approved reimbursement for annual screening using chest CT for smokers
- This was based on promising results for the value of annual CT screening from the National Lung Cancer Screening Trial
- But what if we could perform personalized screening not just based on age and smoking history but based on index of likelihood of developing cancer or other diseases?
- If so, what databases might be useful to guide our screening?

PLCO

- Published in 2009, the PLCO Screening Trial enrolled ~155,000 participants to determine whether certain screening exams reduced mortality from prostate, lung, colorectal and ovarian cancer
- The Prostate, Lung, Colorectal and Ovarian Cancer (PLCO) Screening Trial dataset provides an unparalleled resource for matching patients with the outcomes of demographically or diagnostically comparable patients
- These matched data can be used to inform a more sophisticated, personalized diagnostic decision-making process by tailoring imaging and testing follow-up intervals or even guiding intervention and prognosis
- They can also be incorporated into CAD algorithms to improve diagnostic efficacy by provided a priori likelihood of disease information.

PLCO Dataset

Table 2. Modified Logistic-Regression Prediction Model (PLCO_{M2012}) of Cancer Risk for 36,286 Control Participants Who Had Ever Smoked.*

Variable	Odds Ratio (95% CI)	P Value	Beta Coefficient
Age, per 1-yr increase†	1.081 (1.057–1.105)	<0.001	0.0778868
Race or ethnic group‡			
White	1.000		Reference group
Black	1.484 (1.083–2.033)	0.01	0.3944778
Hispanic	0.475 (0.195–1.160)	0.10	–0.7434744
Asian	0.627 (0.332–1.185)	0.15	–0.466585
American Indian or Alaskan Native	1		0
Native Hawaiian or Pacific Islander	2.793 (0.992–7.862)	0.05	1.027152
Education, per increase of 1 level†§	0.922 (0.874–0.972)	0.003	–0.0812744
Body-mass index, per 1-unit increase†	0.973 (0.955–0.991)	0.003	–0.0274194
Chronic obstructive pulmonary disease (yes vs. no)	1.427 (1.162–1.751)	0.001	0.3553063
Personal history of cancer (yes vs. no)	1.582 (1.172–2.128)	0.003	0.4589971
Family history of lung cancer (yes vs. no)	1.799 (1.471–2.200)	<0.001	0.587185
Smoking status (current vs. former)	1.297 (1.047–1.605)	0.02	0.2597431
Smoking intensity¶			–1.822606
Duration of smoking, per 1-yr increase†	1.032 (1.014–1.051)	0.001	0.0317321
Smoking quit time, per 1-yr increase†	0.970 (0.950–0.990)	0.003	–0.0308572
Model constant			–4.532506

* To calculate the 6-year probability of lung cancer in an individual person with the use of categorical variables, multiply the variable or the level beta coefficient of the variable by 1 if the factor is present and by 0 if it is absent. For continuous

PLCO Demo Cancer Risk in Women with BMI Over 30



Patient Characteristics:

Age: 49 - 78

Height (inches): 48 - 84

Weight (pounds): 70 - 399

Gender:

Male

Female

All

Education:

All | None

Less than 8 years

8-11 years

12 years or completed High School

Post High School training other than College

Some College

College Graduate

Postgraduate

Hispanic:

All | None

Find Cancers

Results returned in 4.236 seconds

Total Matches (experimental): 18427

Total Matches (control): 143136

Cancer Type	Odds Ratio (CI)
Breast	2.07 (1.93-2.22)
Endometrium	1.58 (1.37-1.83)
Lung	0.59 (0.52-0.67)
Colorectum	0.88 (0.77-1.01)
NonHodgkin's Lymphoma	0.9 (0.75-1.08)
Melanoma	0.64 (0.53-0.78)
Kidney and Renal Pelvis	1.05 (0.85-1.3)
Ovary	0.92 (0.73-1.16)
Leukemia	0.71 (0.56-0.9)
Pancreas	0.77 (0.6-0.99)
Bladder	0.28 (0.21-0.37)
Thyroid	1.54 (1.13-2.11)

PLCO Cancer Search Tool

PLCO Cancer Search

Patient Characteristics:

Age: 49 - 78

Height (inches): 48 - 84

Weight (pounds): 70 - 399

Gender:

- Male
- Female
- All

Education:

All | None

- Less than 8 years
- 8-11 years
- 12 years or completed High School
- Post High School training other than College
- Some College
- College Graduate
- Postgraduate

Control Exp



Baseline Cohort

141444 matches



Experimental Cohort 1

141444 matches

+ New cohort

Analyze

Results returned in 5.022 seconds

Total Matches (experimental): 141444

Total Matches (overall): 141444

Cancer Type	Odds Ratio (95% CI)	Experimental Rate (cases/total)	Overall Rate (cases/total)
Prostate	1 (0.97-1.03)	11.22% (7834/69817)	11.22% (7834/69817)
Breast	1 (0.96-1.04)	2.97% (4200/141444)	2.97% (4200/141444)
Lung	1 (0.95-1.05)	2.34% (3305/141444)	2.34% (3305/141444)
Colorectum	1 (0.94-1.06)	1.49% (2108/141444)	1.49% (2108/141444)

Current Smokers PLCO Risk

Gender:

- Male
- Female
- All

Education:

- All | None
- Less than 8 years
 - 8-11 years
 - 12 years or completed High School
 - Post High School training other than College
 - Some College
 - College Graduate
 - Postgraduate

Marital:

- All | None
- Married or living as married
 - Widowed
 - Divorced
 - Separated
 - Never Married

Occupation:

- All | None
- Homemaker
 - Working
 - Unemployed

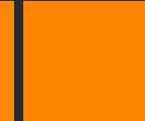
Results returned in 3.363 seconds

Total Matches (experimental): 15255

Total Matches (overall): 141444

Cancer Type	Odds Ratio (95% CI)	Experimental Rate (cases/total)	Overall Rate (cases/total)
Lung	4.18 (3.92-4.46)	9.09% (1386/15255)	2.34% (3305/141444)
Prostate	0.75 (0.69-0.81)	8.61% (709/8238)	11.22% (7834/69817)
Breast	0.8 (0.72-0.89)	2.39% (364/15255)	2.97% (4200/141444)
Colorectum	1.16 (1.02-1.32)	1.72% (263/15255)	1.49% (2108/141444)
Bladder	1.75 (1.53-2)	1.65% (252/15255)	0.95% (1346/141444)
Pancreas	1.57 (1.29-1.92)	0.74% (113/15255)	0.47% (669/141444)
NonHodgkin's Lymphoma	0.83 (0.68-1.02)	0.68% (104/15255)	0.82% (1164/141444)
Kidney and Renal Pelvis	1.22 (0.98-1.51)	0.62% (95/15255)	0.51% (723/141444)
Larynx	4.27 (3.32-5.49)	0.58% (89/15255)	0.14% (194/141444)
Melanoma	0.57 (0.46-0.71)	0.55% (84/15255)	0.95% (1350/141444)
Lip, Oral Cavity, Pharynx	2.53 (1.98-3.23)	0.54% (82/15255)	0.21% (302/141444)
Leukemia	0.9 (0.71-1.14)	0.51% (78/15255)	0.57% (800/141444)
Not Ascertained	2.02 (1.54-2.64)	0.43% (65/15255)	0.21% (299/141444)
Stomach	1.79 (1.35-2.37)	0.39% (59/15255)	0.22% (306/141444)

“Former Smokers” PLCO Risk 124



All

Education:

All | None

- Less than 8 years
- 8-11 years
- 12 years or completed High School
- Post High School training other than College
- Some College
- College Graduate
- Postgraduate

Marital:

All | None

- Married or living as married
- Widowed
- Divorced
- Separated
- Never Married

Occupation:

All | None

- Homemaker
- Working
- Unemployed
- Retired
- Extended Sick Leave
- Disabled

Cancer Type	Odds Ratio (95% CI)	Experimental Rate (cases/total)	Overall Rate (cases/total)
Prostate	0.96 (0.92-1)	10.85% (3938/36306)	11.22% (7834/69817)
Lung	1.16 (1.09-1.23)	2.71% (1666/61446)	2.34% (3305/141444)
Breast	0.86 (0.81-0.91)	2.56% (1575/61446)	2.97% (4200/141444)
Colorectum	1.07 (0.99-1.15)	1.59% (978/61446)	1.49% (2108/141444)
Bladder	1.36 (1.25-1.49)	1.29% (792/61446)	0.95% (1346/141444)
Melanoma	1.06 (0.96-1.17)	1.01% (619/61446)	0.95% (1350/141444)
NonHodgkin's Lymphoma	0.97 (0.87-1.08)	0.8% (492/61446)	0.82% (1164/141444)
Leukemia	1.1 (0.97-1.24)	0.62% (382/61446)	0.57% (800/141444)
Kidney and Renal Pelvis	1.11 (0.98-1.26)	0.57% (348/61446)	0.51% (723/141444)
Pancreas	1.02 (0.89-1.17)	0.48% (296/61446)	0.47% (669/141444)
Endometrium	0.99 (0.85-1.15)	0.88% (221/25140)	0.89% (635/71627)
Multiple Myeloma	1.12 (0.94-1.34)	0.29% (179/61446)	0.26% (368/141444)
Stomach	1.17 (0.96-1.42)	0.25% (155/61446)	0.22% (306/141444)
Esophagus	1.35 (1.1-1.65)	0.24% (148/61446)	0.18% (252/141444)
Not Ascertained	1.09 (0.89-1.33)	0.23% (141/61446)	0.21% (299/141444)
Lip, Oral Cavity, Pharynx	1.07 (0.88-1.31)	0.23% (140/61446)	0.21% (302/141444)

Evaluation of the Lung Cancer Risks at Which to Screen Ever- and Never-Smokers: Screening Rules Applied to the PLCO and NLST Cohorts

Martin C. Tammemägi^{1*}, Timothy R. Church², William G. Hocking³, Gerard A. Silvestri⁴, Paul A. Kvale⁵, Thomas L. Riley⁶, John Commins⁶, Christine D. Berg⁷

1 Department of Health Sciences, Brock University, St. Catharines, Ontario, Canada, **2** School of Public Health, University of Minnesota, Minneapolis, Minnesota, United States of America, **3** Marshfield Clinic, Marshfield, Wisconsin, United States of America, **4** Pulmonary and Critical Care Medicine, Medical University of South Carolina, Charleston, South Carolina, United States of America, **5** Pulmonary and Critical Care Medicine, Henry Ford Health System, Detroit, Michigan, United States of America, **6** Information Management Systems, Rockville, Maryland, United States of America, **7** Department of Radiation Oncology and Molecular Radiation Sciences, Johns Hopkins Medicine, Baltimore, Maryland, United States of America

Abstract

Background: Lung cancer risks at which individuals should be screened with computed tomography (CT) for lung cancer are undecided. This study's objectives are to identify a risk threshold for selecting individuals for screening, to compare its efficiency with the U.S. Preventive Services Task Force (USPSTF) criteria for identifying screenees, and to determine whether never-smokers should be screened. Lung cancer risks are compared between smokers aged 55–64 and ≥ 65 –80 y.

Methods and Findings: Applying the $PLCO_{m2012}$ model, a model based on 6-y lung cancer incidence, we identified the risk threshold above which National Lung Screening Trial (NLST, $n = 53,452$) CT arm lung cancer mortality rates were consistently lower than rates in the chest X-ray (CXR) arm. We evaluated the USPSTF and $PLCO_{m2012}$ risk criteria in intervention arm (CXR) smokers ($n = 37,327$) of the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial (PLCO). The numbers of smokers selected for screening, and the sensitivities, specificities, and positive predictive values (PPVs) for identifying lung cancers were assessed. A modified model ($PLCO_{all2014}$) evaluated risks in never-smokers. At $PLCO_{m2012}$ risk ≥ 0.0151 , the 65th percentile of risk, the NLST CT arm mortality rates are consistently below the CXR arm's rates. The number needed to screen to prevent one lung cancer death in the 65th to 100th percentile risk group is 255 (95% CI 143 to 1,184), and in the 30th to <65th percentile risk group is 963 (95% CI 291 to ∞); the number needed to screen could not be estimated in the <30th percentile risk group because of absence of lung cancer deaths. When applied to PLCO intervention arm smokers, compared to the USPSTF criteria, the $PLCO_{m2012}$ risk ≥ 0.0151 threshold selected 8.8% fewer individuals for screening ($p < 0.001$) but identified 12.4% more lung cancers (sensitivity 80.1% [95% CI 76.8%–83.0%] versus 71.2% [95% CI 67.6%–74.6%], $p < 0.001$), had fewer false-positives (specificity 66.2% [95% CI 65.7%–66.7%] versus 62.7% [95% CI 62.2%–63.1%], $p < 0.001$), and had higher PPV (4.2% [95% CI 3.9%–4.6%] versus 3.4% [95% CI 3.1%–3.7%], $p < 0.001$). In total, 26% of individuals selected for screening based on USPSTF criteria had risks below the threshold $PLCO_{m2012}$ risk ≥ 0.0151 . Of PLCO former smokers with quit time > 15 y, 8.5% had $PLCO_{m2012}$ risk ≥ 0.0151 . None of 65,711 PLCO never-smokers had $PLCO_{m2012}$ risk ≥ 0.0151 . Risks and lung cancers were significantly greater in PLCO smokers aged ≥ 65 –80 y than in those aged 55–64 y. This study omitted cost-effectiveness analysis.

Conclusions: The USPSTF criteria for CT screening include some low-risk individuals and exclude some high-risk individuals. Use of the $PLCO_{m2012}$ risk ≥ 0.0151 criterion can improve screening efficiency. Currently, never-smokers should not be screened. Smokers aged ≥ 65 –80 y are a high-risk group who may benefit from screening.

Deriving Recommendations for Lung Cancer Screening

- Selection of individuals for LDCT lung cancer screening programs using the PLCOm2012 risk ≥ 0.0151 criterion should improve screening efficiency compared to selection by USPSTF criteria
- Currently, never-smokers should not be screened
- Lung cancer screening of high-risk older smokers (65–80 y) should be encouraged.



Creating Local/Regional Databases from Clinical Data

- Would also like to be able to collect data at the University of Maryland, within the Department of Veterans Affairs Hospitals in Maryland and then nationally that could establish a similar database
- Then could provide report that gave reference database such as NLST with likelihood of malignancy and also gave local reference to a specific population and then taking into account PLCO data

NLST and PLCO Next Steps

- Huge implications for screening, e.g. reduce cost from \$240,000 for smokers over 50 years old to a lower cost for a higher risk cohort for CMS concerns
- Has major implications for Bayesian pre-test probability data to assist in diagnosis
- Working with multiple vendors, demonstrating ability to incorporate this into the workflow with ability to “click” on nodule and then have automated lesion characterization, lookup from EMR and then access reference database “service” to get information about likelihood of malignancy
- Would like to incorporate these data into routine applications such as CAD software that could take a priori characteristics to help to CAD_x in addition to current CAD_e

Next Steps:

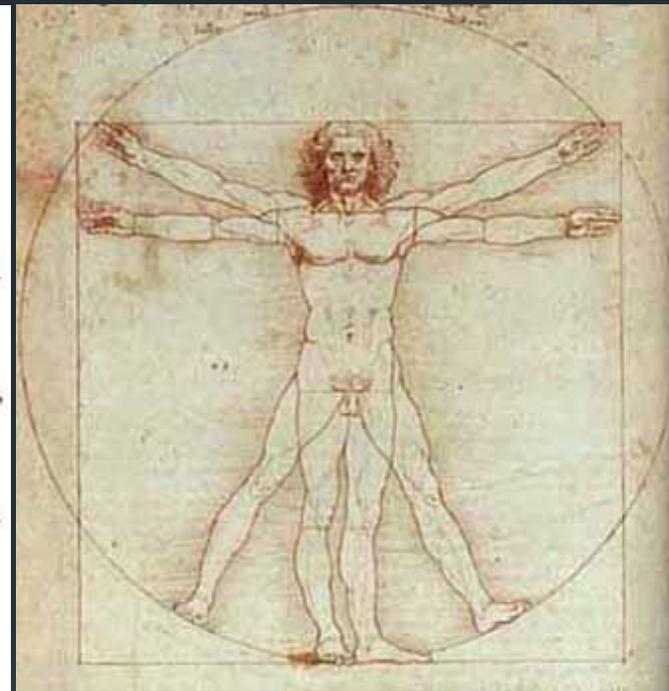
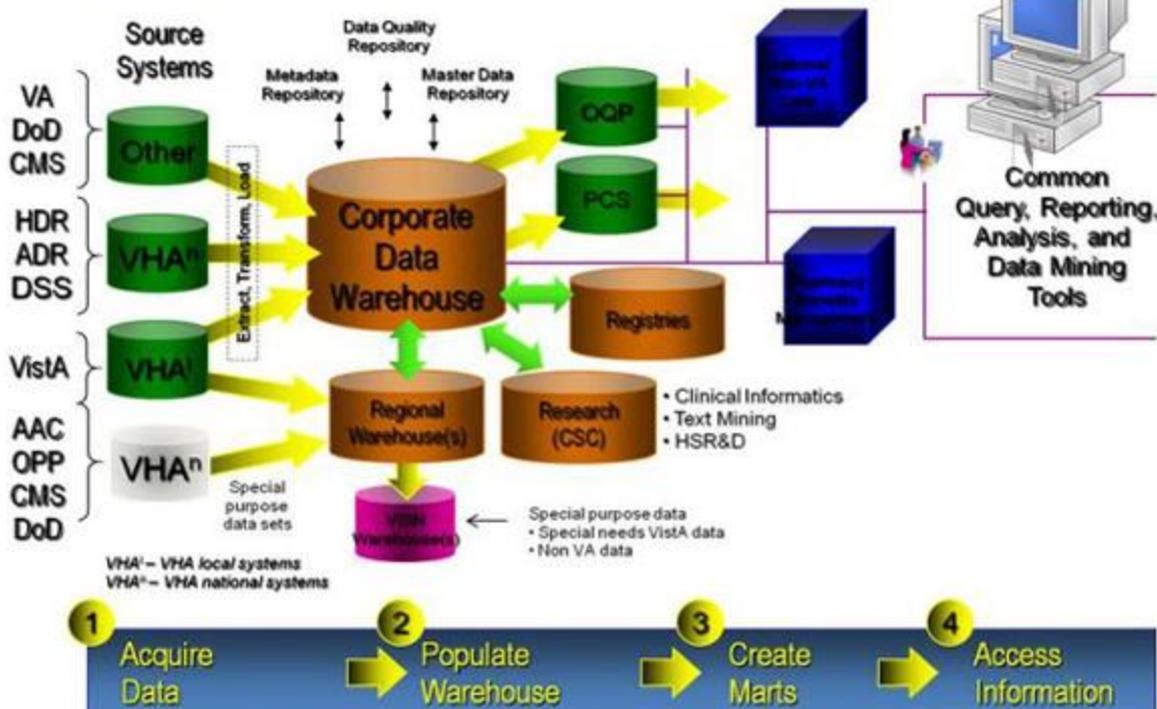
- NLST and PLCO are just examples and just scratch the surface of the incredible amount of data that is out there to be mined for help in patient safety and decision support:
 - Synthesis of EMR
 - Patient safety checker like spell checker
 - Assistance in diagnosis and treatment recommendations

Tackling Truly Huge Clinical Databases

VA's Corporate Data Warehouse Vinci

Combing with Imaging for Cross Correlations e.g. Coronary Artery Calcification, Aortic Size, Lung Texture, Renal Artery Disease etc.

VHA Data Warehousing Visual Architecture – Current Vision



Challenges with Mining Truly Huge Clinical Databases such as VINCI Database With 32 Million Patients over 17 Years

- Issues about how to mine such a large database from a speed/sampling perspective
- Do we shoot for most common treatment suggestion, or one from experts, or one that empirically seemed to work best but then we wouldn't be able to push the envelope toward new treatments/ideas
- If we personalize for specific patient, how do we test that this has been successful?
- Issues about how to debug software and test software that is complicated enough to make decision support questions for medical diagnosis and treatment

- The use of Big data in radiology for clinical applications is still in its infancy
- We cannot use the same statistical approaches to discover and mine data from high dimensional databases
- Medical imaging has over the years tackled high dimensional data in medical images and could be well poised to help out with challenges in mining data from the EMR

Conclusion

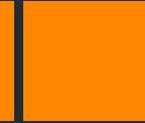
- In order to accomplish this major leap forward we need
 - Access to existing **datasets** from clinical trials such as NLST and our own local EMR databases to achieve personalized, precision medicine
 - Re-think CAD_e and CAD_x and how it is delivered:
 - Eliminate its black box nature and have it utilize standardized databases to develop algorithms
 - Have CAD indicate its level for suspicion of disease and be able to drill into the factors that it used in its decision
 - Make CAD interactive with the radiologist rather than its current role as a second reader

Conclusion

- We need to promote general adoption by vendors of image tagging “standards” such as AIM (annotation and image mark-up)
- Make these “tags” available to EMR algorithms that are used to make diagnostic and therapeutic decisions such as whether to put a patient on statins

Conclusion

- We in Diagnostic Imaging are on the cusp of exciting new opportunity to use “Big Data” for real time decision support for screening criteria, clinical diagnosis and treatment
- However we are already seeing applications that could change the way in which medicine and specifically radiology is practiced in the future
- I believe that experts in medical imaging can help guide the way toward better and safer patient care



Practical Applications and Pitfalls Of 'Big Data' For Decision Support In Medical Imaging and Informatics

Can Medical Imaging Community Play a Significant Role?

Eliot Siegel, MD, FACR, FSIIM

Professor and Vice Chair University of Maryland School of Medicine
Department of Diagnostic Radiology

Professor Computer Science University of Maryland Baltimore County

Professor Biomedical Engineering University of Maryland College Park