Informatics

For a Practicing Cytopathologist

Victor Brodsky, MD
Medical Director of Informatics

The Outline

What is Informatics?
Why do I need to know about it?
How did we get here?
What do I need to know about electronics?
What are the basics of computer science?
What constitutes fault tolerance?
What do I need to know about hacking?
Should we "just buy IT"?
Some useful stats for whole slide imaging

Weill Cornell/NYP Hospital

- 824 Beds
- 35 Faculty Pathologists (AP + CP)
- 41,000 Surgical specimens / year
- 54,000 Cytopathology specimens / year
  - 42,000 Gyn
  - 12,000 Non-Gyn
  - 4,500 FNAs
- 22 Residents
- 9+ AP Fellows

Definition

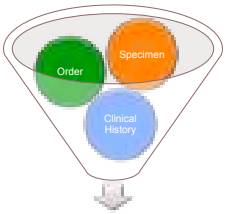
According to Merriam Webster:

"the collection, classification, storage, retrieval, and dissemination of recorded knowledge treated both as a pure and as an applied science"

I prefer: "innovative and progressive information management via application of technology"

Pathology Laboratory

An information factory



Test Results

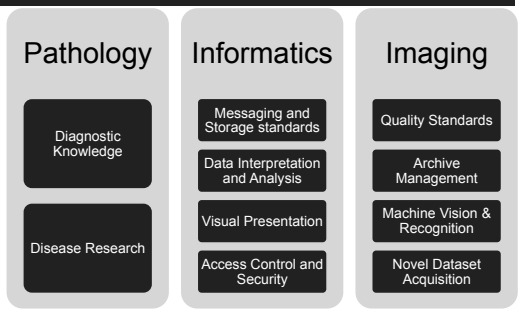
Growing amount of data

Over 3000 clinical tests are currently offered by reference laboratories (3.5 mil/year here)

Additional experimental tests are done for research (and personal genomes are coming)

As an extreme example, a reference laboratory can perform about 20 million tests per year, employing over 3000 people

Domains of Expertise



Informatics: Operational

- Are the digital requisition forms optimal and effective?
- Are the reports getting delivered to all recipient systems?
- Orders interface vs Results interface
- Body site vs Specimen Type vs Specimen Source
- Laboratory test naming and mapping
- Is the coolscope accessible?

Informatics: Operational

- Barcoding and Tracking of containers
  - Location Awareness
  - Productivity monitoring
  - Proactive alerts
- Improving interfaces between systems
  - Eliminating paper: errors, duplicate data entry
- Implementing voice recognition for dictations
- Moving to discrete data element capture
- Streamlining ICD10/CPT coding of the reports

Informatics: Regulatory

- HIPAA Compliance
- FERPA Compliance
- HITRUST Compliance
- CLIA Compliance
- Other Federal & State Law Compliance
- CAP Inspection Compliance
- Hospital policy compliance
- Data leakage protection

Informatics: Innovation

- Keeping up to date
- Vendor selection
- Vendor negotiation
- Vendor collaboration
- Product development by the vendor
- Implementation of transformative technology
  - Speech Recognition
  - Barcoding and Tracking
  - Whole slide scanning

Informatics: Research

- Supporting researchers
  - Data analysis
  - Hardware availability
- Collaborating with researchers
  - Image analysis
- Informatics Research
  - Innovative: Experience Profiling
  - Basic science: Protein folding
  - Measuring impact on quality of care

History

1000 BC Babylonians develop first abacus
1729 Stephen Gray discovers electrical conduction.
1833 Charles Babbage creates first plans for a logical calculating computer
1837 Samuel Morse invents the first telegraph

History

1934 Paul Otlet generally envisions a global network of computers, or "electric telescopes"
1946 ENIAC: first electronic computer
1957 IBM debuts first dot-matrix printer
1959 Robert Ledley, D.D.S. (the inventor of the CT Scanner) and Lee B. Lusted publish…

History

1959


SCIENCE

Reasoning Foundations of Medical Diagnosis

History

1960 Robert S. Ledley and Lee B. Lusted, "Computers in Medical Data Processing" paper
1962 J.C.R. Licklider at MIT envisions a worldwide network of computers.
1962 William R. Best proclaims in JAMA: "If the computer is going to do so much for medicine, why must we wait?"

History

1965 Robert Ledley's book "Use of Computers in Biology and Medicine" is published
1967 HELP (Health Evaluation Through Logical Processing) EMR launched at LDS hospital (IHC) in Utah
1968 COSTAR (Computer Stored Ambulatory Record) at MGH; 30,000 patients stored by 1973

History

1968 MHTS (Multiphasic Health Testing System) launched at Kaiser Permanente in San Francisco
1968 Larry Weed describes POMR
1969 US Department of Defense develops ARPANET, the first packet switching computer network and the progenitor of the Internet.

History

- 1971 IBM demonstrates the first speech recognition software.
- 1971 MEDLINE (MEDLARS Online) is launched by US National Library of Medicine (Precursor of PubMed)
- 1973 Regenstrief Institute launches their first electronic Medical Records System, RMRS

History

- 1976 Clinical Computing System is launched at Beth Israel Hospital in Boston
- 1981 IBM PC is introduced
- 1981 FujiFilm digitizes X-Ray images
- 1989 The first Bioinformatics degree program (undergraduate) is opened at Carnegie Mellon University

History

- 1990 Human Genome Project launched
- 1991 Institute of Medicine recommends that by the year 2000 every physician should be using computers in their practice
- 1996 PubMed makes Medline searching available on the internet for free
- 1996 HIPAA

History

- 1997 Deep Blue wins the Chess match
- 1997 Trac Cell 2000 approved by FDA
- 1998 AutoPap 300 for primary screening
- 2003 Human Genome Project Completed
- 2008 GINA is passed
- 2008 Over 200 EMR vendors to choose from are available
- 2011 Watson wins Jeopardy!

Talk to IT

Ride IT

Navia – first fully self-driving car for sale now

Source: http://induct-technology.com/en/products/navia

See IT

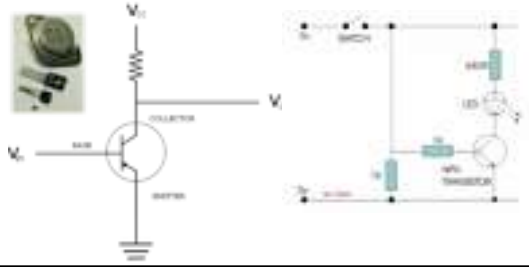
Google Glass
Myo
Oculus Rift
Sony Morpheus
Atlas / Petman

Image Source: http://www.sony.com/SCA/company-news/press-releases/sony-computer-entertainment-america-inc/2014/sony-computer-entertainment-announces-project-morp.shtml

Say Hi to Atlas

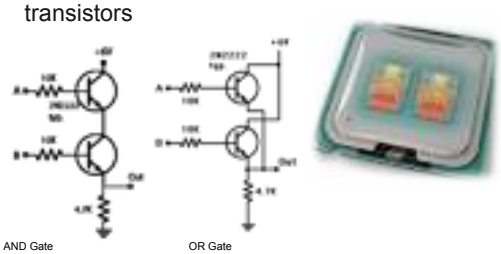
Electronics Course 1/6

Transistor is a “switch”



Electronics Course 2/6

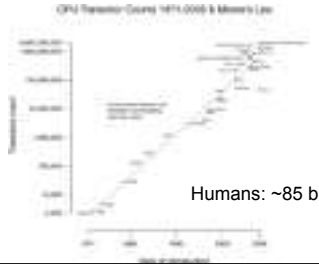
Processors crunch data and are made of transistors



AND Gate OR Gate

Electronics Course 3/6

Transistors: CPU: 2.5 billion GPU: 8 billion



Humans: ~85 billion neurons

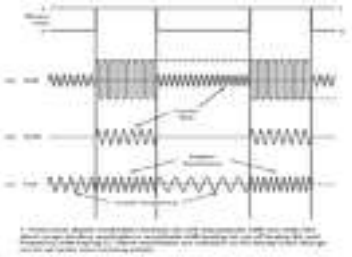
Electronics Course 4/6

Data is stored in memory

- “permanent” = hard drive, flash memory, disc
- “volatile” (zeros out on power off; leaky bucket) = RAM

Electronics Course 5/6

Data is sent via modulated frequencies

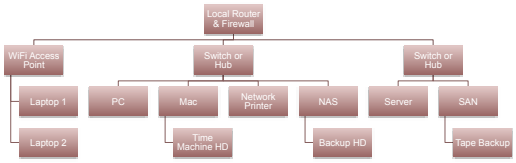


Hertz = 1 cycle/second

Gigahertz = 1 billion cycles / s

Electronics Course 4/4

Simplified Ethernet Network Diagram



Computer Science: 1/3

Bit = Basic unit of information, a 1 or a 0

Since there are only two possible states (nicely corresponding to either presence or absence of electrical current) representation of information in bits is called binary code

1 bit can represent $2^1 = 2$ concepts 0 - I

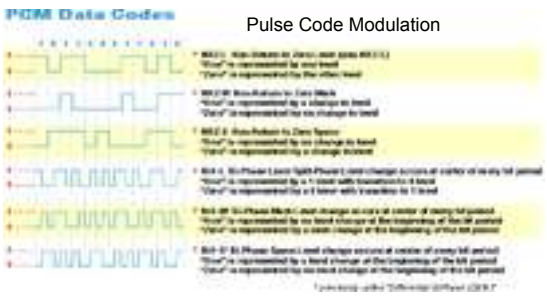
2 bits can represent $2^2 = 4$ concepts 00 - 0I - I0 - II

3 bits can represent $2^3 = 8$ concepts

Byte = 8 bits (ASCII http://www.ascii-code.com/)

24 bits can represent a color in RGB

Electronic Course BONUS



Pulse Code Modulation

Computer Science: 2/3

A processor expects 2 types of bits:

"substrate" data ("Video")

"instructions" on what to do with that data ("Player")

A processor understands "machine code"

Humans write instructions in programming languages (Assembler, C, Java…) and then use a "compiler" program to translate them to "machine code" (http://rosettacode.org/wiki/Hello_world/Text)

A "file" is a set of bits, usually containing either executable instructions or substrate (or both!)

Computer Science 3/3

A "Database" is a collection of related data values

SQL is a language to write to and read from SQL database tables

There are also many other database types (Graph Databases, "NoSQL" Databases…)

To take full advantage of multi-core & multi-processor systems, multi-threaded code must be written; GPUs can be faster than CPUs

Fault Tolerance

Electrical Power
Physical Access Security
Redundant Memory (Hard drives): RAID
Redundant Servers
ECC RAM
Redundant Network Hardware
Backup
Monitoring

Hacking

Social Engineering (Who is speaking to you?)
Vulnerabilities + Malformed requests:
Heartbleed: http://xkcd.com/1354/
SQL Injection:
More? Metasploit

Keep software updated!

Life without standards

- Manual workarounds
  - Print from A, scan or retype into B
  - Repetitive data entry by hand, paperwork
- Consulting fees for integration
  - Redone with system upgrades
  - Interfaces have yearly maintenance fees
- Difficult to search unstructured data
- Vendor lock-in
- Bottom line: Inconvenient and expensive!

Types of Standards

Value Storage formats
CSV: 1, 2, 3, 4, 5
JSON: { "A": "1", "B": "2" }
XML: Hi
RDF: :sky :isOfColor :blue
OWL / DMCI (Dublin Core)
http://obofoundry.org/
Data compression
JPEG – "lossy"
ZIP – "lossless"

Types of Standards

Domain Knowledge
AJCC – Cancer Staging
CAP Cancer Checklists
Controlled vocabularies
ICD9 / ICD10
CPT
SNOMED-CT - Systemized Nomenclature for Medicine
LOINC

Types of Standards

Report Content Structure
CDA Implementation Guide
Layout
PDF
HTML / CSS; ePub
Messaging between equipment or software
DICOM – Digital Imaging and Communication in Medicine
HL7 – Health Level 7
NAACR Volume V – North American Association of Central Cancer Registries

Future Standards

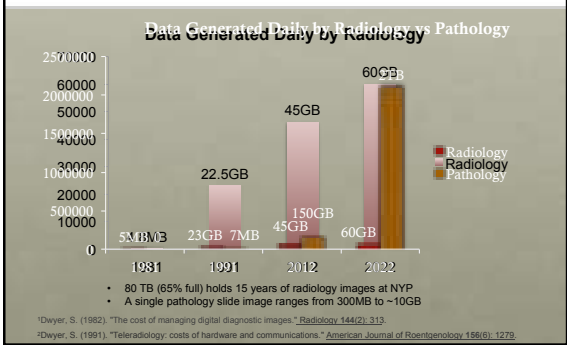
- Form Rules
- Specimen Types
- Location within the human body
- Value Set transfer
- Ontology Map storage (RIF?)
- Value Transformation (units, normal ranges)
- Change Notification
- Versioning queries and storage
- Hanging Protocols
- Naming Conventions for tests
- Barcodes

Fun With Standards

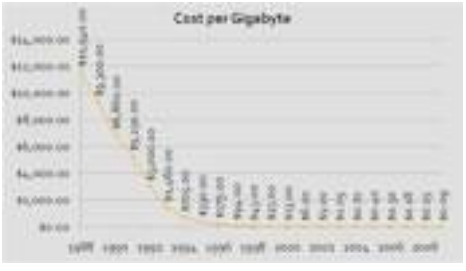
Goals of Scanning Slides

1. Prepare pathologists and residents for the future
2. Provide and request consultations globally
3. Stop glass slides from disappearing
4. Make productivity and quality measurable
5. Match report with images
6. Highlight unviewed areas
7. Display two images side by side or overlay them
8. **Maintain a copy of slides sent/received for consultation**
9. Laser micro dissection for DNA, RNA, and protein retrieval

Medical Image Sizes



Price of Storage



Data till 2002 from: National Science Foundation, Division of Science Resources Statistics, Science and Engineering Indicators–2002, Arlington, VA (NSB 02-01) [April 2002]

Assessing your needs

- Purpose
- Volume per day
- Immunoflourescence scanning
- Magnification higher than ~40X
- Larger-than-usual glass slides

What you will need

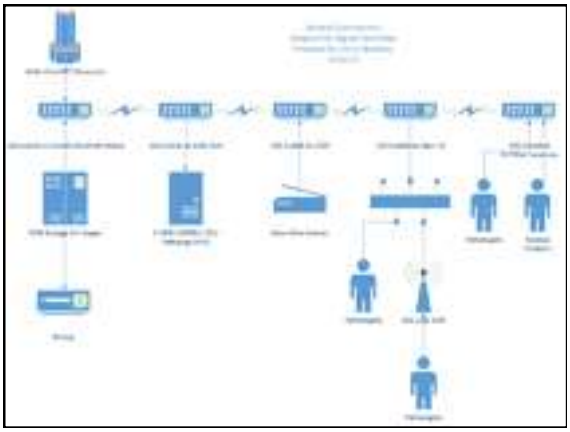
1. A room (or at least a sizable corner)
2. Reliable electricity (an outlet with backup or a local UPS)
3. Two human operators (or at least a part of an FTE)
4. A network jack (10GbE or at least 1GbE)
5. A server to run your web-based "photo album" PACS (3 of them: Live, Dev, Test)

What you will need

6. Expandable storage for your images (SAN or a RAID array)
7. A HIPAA-compliant server room to host your server & storage
8. A fraction of your FTE to be responsible for administering the server & storage
9. Expandable storage to serve as backup (ideally geographically independent)
10. Routers & switches between your scanner, server, storage, backup (10GbE ? Overloaded?)

What you will need

11. 1GbE+ network cables & routers+switches to reach all viewing stations.
12. Big, calibrated "true color" monitors (4K, 28”+)
13. Video cards that can handle those 4K monitors
14. PCs that can handle those video cards
15. LDAP/AD Integration to enable users to log in with their institutional accounts



What you will need

16. Scanner that can read your barcodes on the slides (what if there is more than one?)
17. Interface with your LIS to allow "click-out" access to associated images & auto-population of data
18. Integration of images into reports and other data
19. FDA Approval (or at least a local study as per CAP guidelines)
20. CLIA license for remote users

What you will need

Pathologists who are willing to give it a chance

What you might need

Image analysis software (& server hardware to run it on)

Resolution

- At “40X”, or a 400X magnifications
- Field of view diameter: 0.45mm
- 18cm (7 inch) diameter (400x0.45mm)
- Human retina needs ~350dpi
- 2480 pixels (7x350dpi) along its diameter
- 2480 x 2480 = 6,150,400 pixels – 6 MP

Current “HD” and “4K”

8,294,400

6,150,400

2,073,600

414,720

Receiving Consultations

Accepting "one-off" consultation images
- Secure patient information and image transfer
- Willing pathologists
- Collecting payment

Becoming a partner
- Avoiding lag (“ping”)
- Ensuring timely response on your end

20X Scans

0.46 microns/pixel
3.44 minutes per slide
5.5% rescan rate (adding 0.19 minutes of scan time per slide)
Manual steps: 1.55 minutes per slide
5.18 minutes per slide total
Average file size = 498 megabytes

40X Scans

0.22 microns/pixel
22.96 minutes per slide
2.45% slides rescan rate (adding 3.07 minutes of scan time per slide)
Manual steps: 1.55 minutes per slide
27.58 minutes per slide total
Average file size = 726 megabytes.

Subsequent analysis

- Average scan time per slide at "20x" was 3.76 min.
  - Factoring in the observed 2.5% rescan rate increased the time to 3.86 minutes.
  - Including the pre and post-scan manual steps, which averaged 1.55 minutes per slide, the total average time per "20x" slide was 5.41 minutes.
- Average scan time per slide at "40x" was 21.10 min.
  - Factoring in the observed 5.5% rescan rate increased the time to 21.50 minutes.
  - Including the manual steps, the total average time per "40x" slide was 24.05 minutes.

Subsequent analysis

- File size
  - Scanning at "20x" resulted in an average file size of 560 megabytes
  - Scanning at "40x" resulted in an average file size of 605 megabytes.
- Throughput
  - An average of 40.55 slides per day were scanned
  - Maximum of 154 slides per day; 3.5 hour FTE

Tissue section

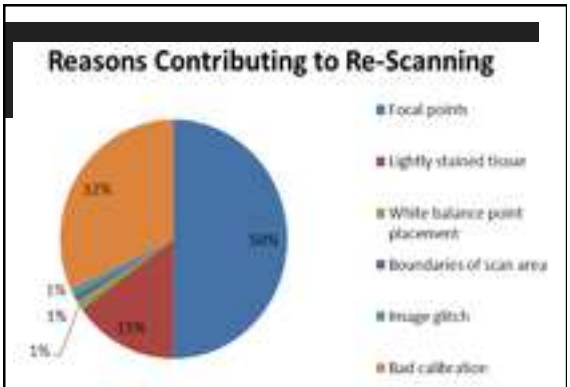
"Industry standard" = 1.5 x 1.5cm = 2.25cm^2

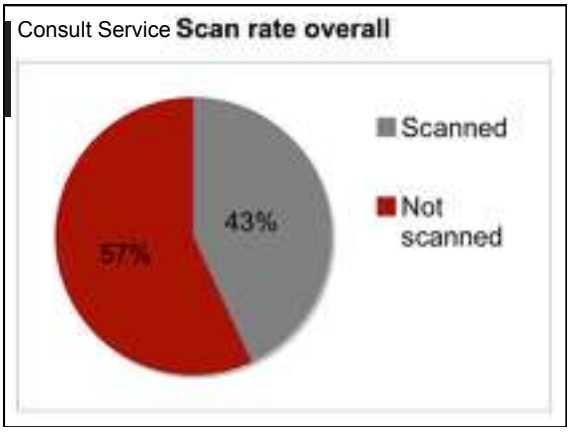
11,736,466,666 pixels = 108,335x108,335

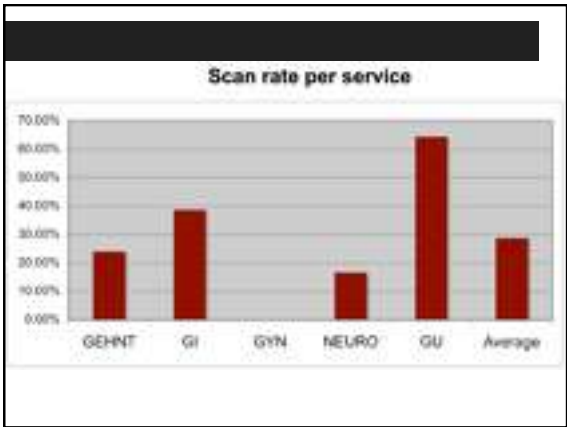
108,335 x 0.22 microns/pixel at "40x" =

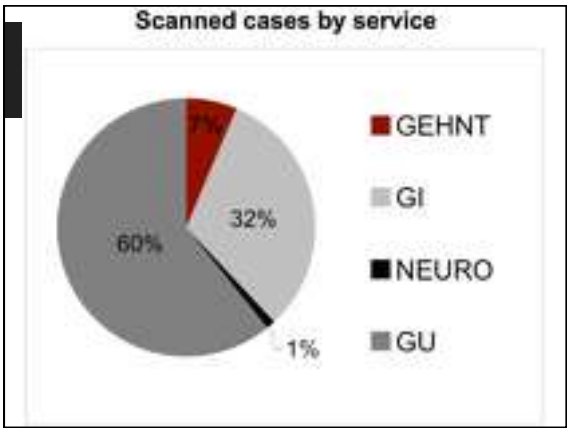
2.38337cm x 2.38337 = 5.68 cm^2

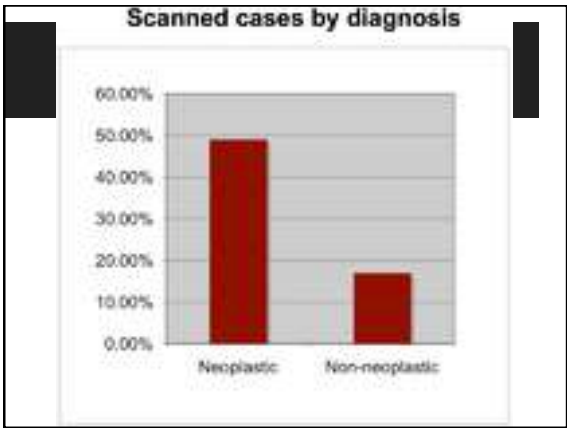
Factor of **2.5246**











Overall Scanner Use

- Out of all consult cases received, 43% were scanned.
-
- The cases contained an average of 9.40 slides each.
- 1 to 12 slides were scanned per case.
- Average 1.98 slides scanned per case.

Per Consult Service

- For the five consult services included in the survey, the percentage of scanned cases ranged from 64% (56/87) to 0% (0/17).
- 49% of cases with neoplastic diagnoses were scanned.
- 17% of non-neoplastic cases were scanned.

Time and Storage

- Total over 8 weeks:
  - 15.71 hours of scan time
  - 90.6 GB of storage
- Weekly average of 1.96 hours of scan time and 11.3 GB of storage (for non-Breast consult cases).
- Scanning only selected slides from each case yielded savings of 58.96 hours of scan time and 340.1 GB of storage over an 8-week period.

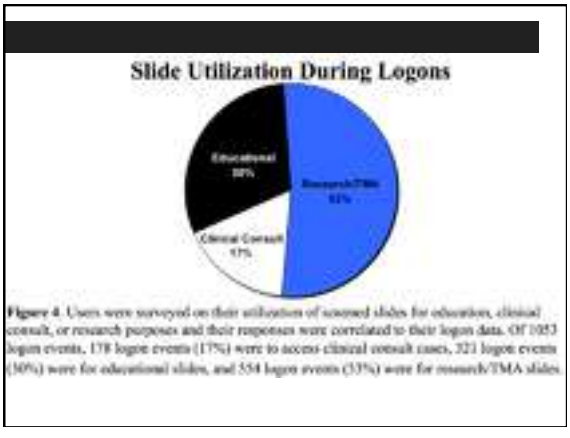


Figure 4. Users were surveyed on their utilization of scanned slides for education, clinical consult, or research purposes and their responses were correlated to their logon data. Of 1053 logon events, 178 logon events (17%) were to access clinical consult cases, 321 logon events (30%) were for educational slides, and 554 logon events (53%) were for research/TMA slides.

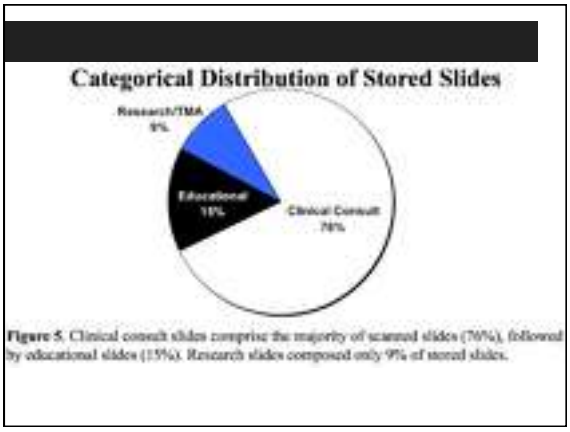
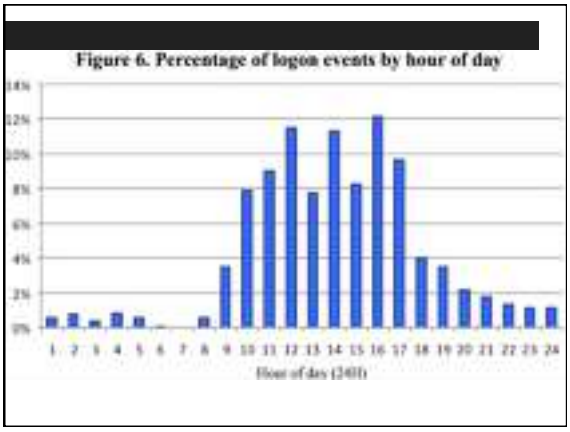
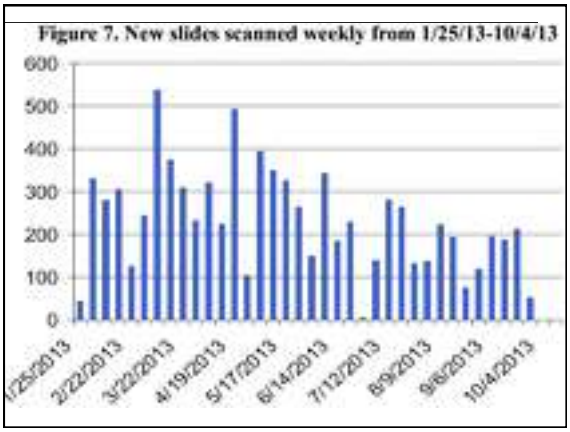
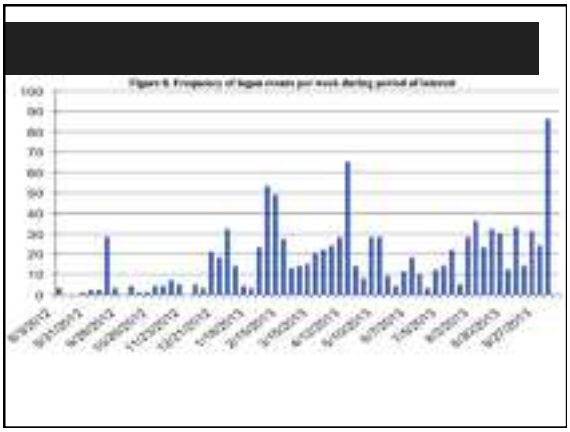


Figure 5. Clinical consult slides comprise the majority of scanned slides (76%), followed by educational slides (15%). Research slides composed only 9% of stored slides.







Imaging - TruePurpose

- 10% to 30% discrepancy between different surgical pathologists among cases where the appearance of tissue warranted obtaining another pathologist's opinion[10]
- Neuropathologists disagree with their own previous diagnosis 25% to 48% of the time[11]
- "Error rates in the order of 30% in radiology plain film reporting have consistently been found."[12]

[10]Gupta D, Layfield LJ. Prevalence of inter-institutional anatomic pathology slide review: a survey of current practice. Am J Surg Pathol. 2000 Feb;24(2):280-4.
[11]Mittler MA, Walter BC, Stopa EG, Observer reliability in histological grading of astrocytoma stereotactic biopsies. J Neurosurg. 1996 Dec;85(6):1091-4.
[12]A.Vohrah, J. C. (2003). "Clinical governance: two years experience of reporting discrepancy review in radiology." Journal of Diagnostic Radiography and Imaging (5): 27–32.

CPath

Systematic Analysis of Breast Cancer Morphology Uncovers Stromal Features Associated with Survival (Beck, AH)

"We applied the C-Path system to microscopic images from two independent cohorts of breast cancer patients [from the Netherlands Cancer Institute (NKI) cohort, n = 248, and the Vancouver General Hospital (VGH) cohort, n = 328]. The prognostic model score generated by our system was strongly associated with overall survival in both the NKI and the VGH cohorts (both log-rank P ≤ 0.001). This association was independent of clinical, pathological, and molecular factors. **Three stromal features were significantly associated with survival, and this association was stronger than the association of survival with epithelial characteristics in the model. These findings implicate stromal morphologic structure as a previously unrecognized prognostic determinant for breast cancer.**" – November 9th, 2011

Distributional Semantics

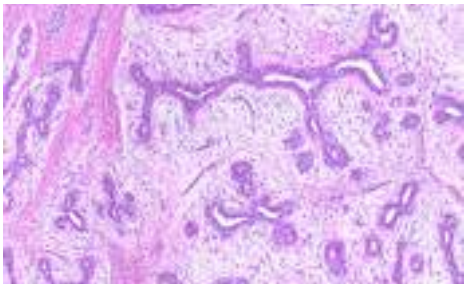
Zellig Harris: Meaning is derived from co-occurrence frequency with other [words]

- 2011 Google launches Search by Image
- 2012 Google Launches Goggles
- 2013 Google launches Photo Search
- 2014 GaussianFace face recognition algorithm outperforms humans

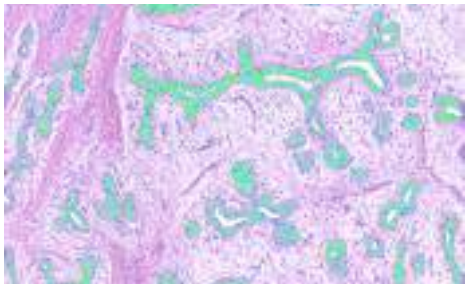
Test your might

Source: http://arxiv.org/abs/1404.3840

One click, unknown image

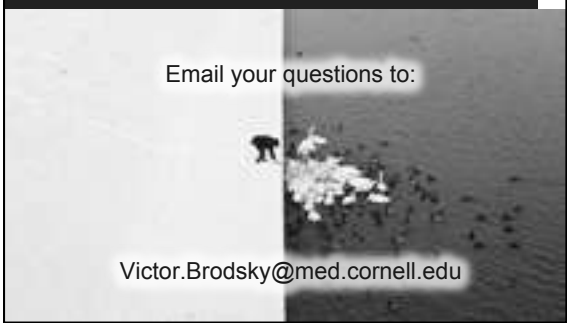


One click, unknown image



New Frontiers

Thank You!



Email your questions to:

Victor.Brodsky@med.cornell.edu