Storage Challenges in the Medical Industry

Nathan Spillers
"The medical field spends 4-7% of revenues on information technology compared to averages of 9-15% in other industries"  

Gartner, 2004
“The most urgent capital needs are buying replacement and new medical equipment, with almost three-quarters of hospital CEOs citing each of these.”

Deloitte and Touche Healthcare Survey, 2005
"[At hospitals,] the single largest spend on software is writing interfaces between proprietary systems. That's insane in my book."

Louis Burns, Manager
Intel’s Digital Health Group
The Big Problem

We are really good at generating data.

We are really bad at effectively using it.
IT is usually viewed as “not delivering” in the medical industry

“Measuring ROI for IT in the medical field increasingly viewed as unattainable goal”

eWeek, December 2005
However, Things are Changing

- Life Sciences is not engineering!
- The knowledge a doctor needs to do their job no longer fits inside their head
- We’re starting to recognize and to build the tools to address this
The Main Challenges

• Ever-increasing volumes
• Unstructured data becoming the rule, not the exception
• Growing requirements for “deep, broad” data utilization and analysis
• We’re changing the way we use our data!
The Volume Problem

• Operational for only three years
• 150,000,000+ images stored to date
• Currently storing approximately 3,000,000 images per week
• Approximately 50 TB of RAID used
• Approximately 145 TB of physical tape storage used
• Currently growing at about a terabyte every 9 days
The “images” are changing
And it’s getting worse...

“The 3-D visualization of the heart’s action over time is critical to cardiology diagnosis and treatment”

“The digital video of my last endoscopy is being used to guide the next one”

“My new lab equipment can generate a terabyte on every diagnostic run!”
Text Annotation

Blade Server

Websphere Application Server

UIMA Pipeline

CDA Parse → Tokenizer → Sentence Detector → Part of Speech Tagger → ...

Websphere Application Server

UIMA Pipeline

CDA Parse → Tokenizer → Sentence Detector → Part of Speech Tagger → ...

Message Queue

Clinical Note

Clinical Note

Clinical Note

DB2

Insert Annotation Data
Text Annotation

Input XML

Anxicillin, 200 mg =>
Amoxicillin, 200 mg

1. type, stage, grade, level, phase, class, para, gnda, alpha, sort, group, category
2. status, post, sp =>
   status, post

Abbrev/ Expansion database

Abbreviation disambiguation
MaxEnt models

POS Model Parser PST

Generalized
Weakness models
NE-recognition models

Subset of UMLS/05AA (NCI, MeSH, ICD-03, SNOMEDCT) (disorders, findings, procedures, anatomical sites)

Roles (Medications) Orange Book

CAS Initializer

Sentence detector
Context Free Tokenizer

Context Sensitive Tokenizer

Lexical Variant Generator

Abbreviation Detection/ Expansion

POS + Parser

Semi-Automatic Interactive Learning Environment for DISORDER, FINDING, MEDICATION

Syntactic adjustment

Dictionary Lookup

SNlW Naive Bayes models

Patient Functional Status Detector/ Classifier

SNlW Naive Bayes classifer (Mayo neare-
synonym cluster IDs)

NP Head Identification

Linguistic Transform

Negation Detector

CAS Consumer

Mayo Clinic Life Sciences System (MCLSS) (DRS)

CAS (common annotation structure)

TissueAnnotation
SentenceAnnotation
AbbrviationAnnnotation
OrdererAnnotation
FindingAnnotation
FunctionalStatusAnn

-Confirmed
-Probable
-History of

Term status Identification

1. N IN NP => NP N
   2. N AUX (to be) ADJ/VBN/VBG =>
   ADJ/VBN/VBG N
Unstructured Information

- Clinical notes
  - Dictated
  - Transcribed
  - Scanned in
- Other “Textual” data
New Utilizations

Link to Clinical Note (Document ID is 361834629)

Chief Complaint/Reason for Visit

The patient comes for followup after being dismissed from the hospital.

History of Present Illness

He had induction chemotherapy for treatment of AML. He had a rather lengthy hospitalization during that time which was complicated by neutropenic fever with some indeterminate lesions in the liver for which he has received moxifloxacin and voriconazole. He also had SIRS/septic shock necessitating intubation and hypotension requiring pressor support. He was thought to have a component of COPD and possibly OSA. He did not require oxygen upon dismissal. There were significant problems with fluid retention, and those are resolving. He now denies nausea and vomiting. He considers his appetite to be poor, but he is really pushing himself to eat--especially in the mornings. He denies fever or chills. He does feel quite weak but is able to get around in his home either with a cane or a walker. He cannot do long distances, but he really is trying to increase his activity. He notices some swelling in his legs and ankles that does seem to be stable and may be resolving a little. His wife is doing the site care of the Hickman catheter. She did have some problems with the Carpuject this weekend and did need to seek help at the Emergency Room. She
The Volume Problem Again

- 18M clinical notes
- ~5k clinical notes (created/revised) per day
- Per clinical note stats (avg):
  - Roughly 60 named entities annotated
    - Drugs
    - Disorders
    - Signs/Symptoms
    - Procedures
    - Anatomy
  - Over 100 Ontology Codes assigned
Unstructured Information
Genomic Microarrays
What does this mean?

- This new way of using the data is causing problems:
  - The filesystem approach to accessing storage is breaking down
  - “Traditional” HSM isn’t addressing the problem any more
  - Complexity and cost are killers
What we need...

- Scalable
- Available
- Feature-rich
- Environment
What we need...

- **S**calable
- **A**vailable
- **F**eature-rich
- **E**nvironment
It's time to address this...

“There’s no part of American life right now that is in more need of imagination and new ideas than health care”

Colin Powell, August 2005
Questions?