Structured vs Unstructured Data

Structured Data

ID	Diagnosis	Value
001	E11.9	140
002	I10	145/90
003	J45.909	Normal

DB

Unstructured Data

Patient presents with chest pain...

History of hypertension and diabetes

Physical exam shows...

NLP

Structured Data

- Predefined fields & formats
- Easily queryable
- Standardized codes (ICD, LOINC)
- Direct database storage
- Machine-readable

Unstructured Data

- Free text clinical notes
- Medical images (X-ray, MRI)
- Scanned documents
- Voice recordings
- Requires NLP for extraction

🔀 Hybrid Documents

Many clinical documents combine structured fields (dates, vital signs) with unstructured narratives (clinical impressions)

📊 Structured Data: Detailed Overview

▶ Definition & Characteristics

Structured data is **highly organized information** that fits neatly into predefined fields and tables. It follows a consistent data model with clearly defined relationships, making it easily searchable and analyzable using standard database queries. In healthcare, structured data enables rapid retrieval, statistical analysis, and integration across different systems.

✓ Consistency
Same format across all records

✓ Queryability
SQL and database operations

✓ Scalability
Efficient storage and retrieval

✓ Interoperability
Easy data exchange between systems

▶ Clinical Examples with Visual Representations

Example 1: Laboratory Results

Patient ID	Test Code	Test Name	Result	Unit	Reference Range	Date
P001234	2345-7	Glucose	142	mg/dL	70-100	2024-11-15
P001234	2093-3	Cholesterol	220	mg/dL	<200	2024-11-15
P001234	718-7	Hemoglobin	13.5	g/dL	12-16	2024-11-15

Key Features: Each lab test has a standardized LOINC code (e.g., 2345-7 for glucose), enabling consistent identification across different healthcare systems. Results are stored in specific data types (numeric values) with defined units, making automated analysis and trending possible.

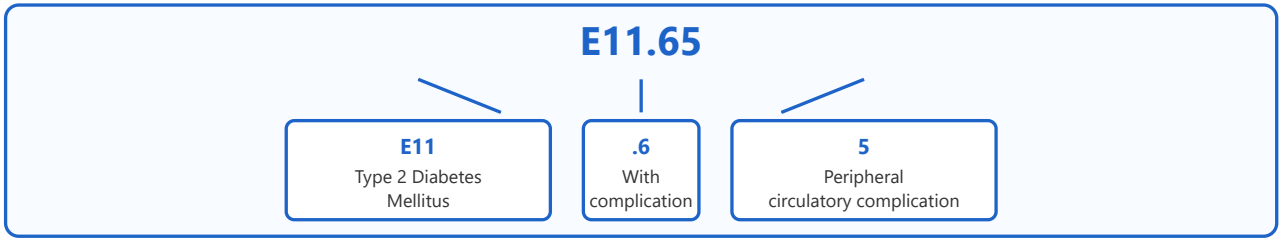
Example 2: Vital Signs Database

Time	BP Systolic	BP Diastolic	Heart Rate	Temp (°F)	SpO2
08:00	120	80	72	98.6	98%
12:00	125	82	75	98.4	97%
16:00	**145**	**92**	78	98.8	98%
20:00	118	78	70	98.6	99%

*** Automated alert triggered for hypertensive reading**

Automation Benefits: Structured vital signs enable automatic alerts when values exceed thresholds (shown in red at 16:00). The system can generate trend graphs, calculate averages, and flag abnormal patterns without human interpretation.

Example 3: ICD-10 Diagnosis Codes



Hierarchical Structure: ICD-10 codes use a hierarchical system where each character adds specificity. This structure enables queries at different levels (all diabetes patients, diabetes with complications, specific complication types) and supports clinical decision support and billing automation.

Example 4: SQL Query Example

```
SELECT patient_id, diagnosis_code, diagnosis_date
FROM diagnoses
WHERE diagnosis_code LIKE 'E11%'
AND diagnosis_date >= '2024-01-01'
ORDER BY diagnosis_date DESC;

-- Returns all Type 2 Diabetes diagnoses from 2024
```

Query Power: Structured data enables precise, rapid queries across millions of records. Healthcare analysts can identify patient cohorts, track disease trends, and generate reports in seconds rather than hours of manual chart review.

▶ Common Healthcare Structured Data Types

- **Demographics:**Patient name, date of birth, gender, address, insurance information
- **Vital Signs:**Blood pressure, heart rate, temperature, respiratory rate, oxygen saturation
- **Laboratory Results:**Blood tests, chemistry panels, microbiology cultures
- **Medications:**Drug names (RxNorm codes), dosages, frequencies, routes of administration

- **Diagnoses:**ICD-10 codes with associated dates and encounter information
- **Procedures:**CPT codes, surgical procedures, interventions
- **Billing Information:**Charges, payments, insurance claims

---

📝 Unstructured Data: Detailed Overview

▶ Definition & Characteristics

Unstructured data is **information without a predefined data model** or organizational structure. It doesn't fit neatly into rows and columns, making it challenging to search, analyze, and process using traditional database methods. In healthcare, unstructured data often contains rich clinical narratives, contextual information, and nuanced observations that structured fields cannot capture. It requires specialized techniques like Natural Language Processing (NLP) and machine learning for analysis.

✓ **Rich Context**

Detailed clinical narratives and observations

✓ **Flexibility**

No rigid format constraints

✓ **Human Language**

Natural expression of clinical thinking

✓ **Multimedia**

Images, audio, video, scanned documents

▶ Clinical Examples with Visual Representations

Example 1: Clinical Progress Note

PROGRESS NOTE - 11/15/2024 14:30

Subjective:

62-year-old male with Type 2 DM presents for follow-up. Patient reports increased thirst and frequent urination over the past 2 weeks. Denies chest pain, shortness of breath, or visual changes. States he has been "stressed at work" and admits to dietary non-compliance.

Objective:

BP 145/92, HR 78, Temp 98.6°F. Patient appears well, slightly anxious. HEENT: PERRLA, no retinopathy noted. Cardiovascular: Regular rhythm, no murmurs. Extremities: Pedal pulses 2+ bilaterally, no edema.

Assessment & Plan:

1. Uncontrolled Type 2 DM - likely due to dietary non-compliance.
   Increase metformin to 1000mg BID. Recheck HbA1c in 6 weeks.
2. HTN - Consider adding ACE inhibitor if BP remains elevated.

NLP Challenges: This note contains crucial information like "dietary non-compliance" and "stressed at work" that provide context for treatment decisions but require sophisticated NLP to extract. Terms like "slightly anxious" contain subjective clinical impressions difficult to quantify.

Example 2: Radiology Report

CHEST X-RAY REPORT

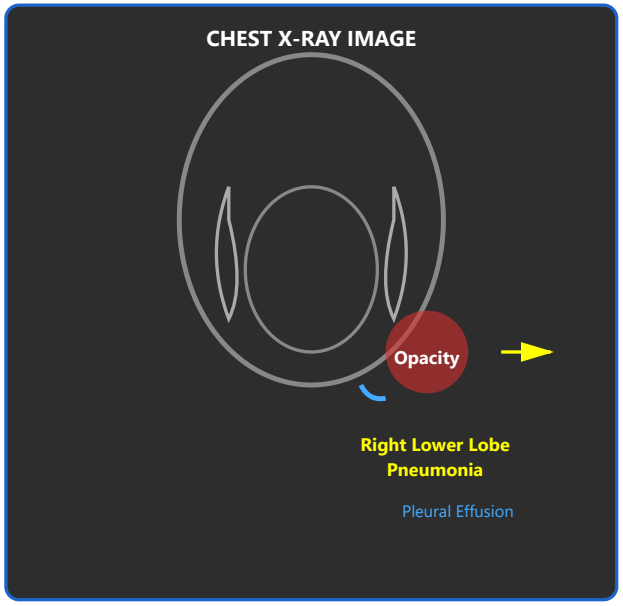
EXAM: Chest PA and Lateral
DATE: November 15, 2024

CLINICAL INDICATION:
Shortness of breath, rule out pneumonia

FINDINGS:
The heart size is normal. There is a focal opacity in the right lower lobe measuring approximately 3 cm, consistent with pneumonia. Small right pleural effusion is noted. No pneumothorax. Osseous structures are unremarkable.

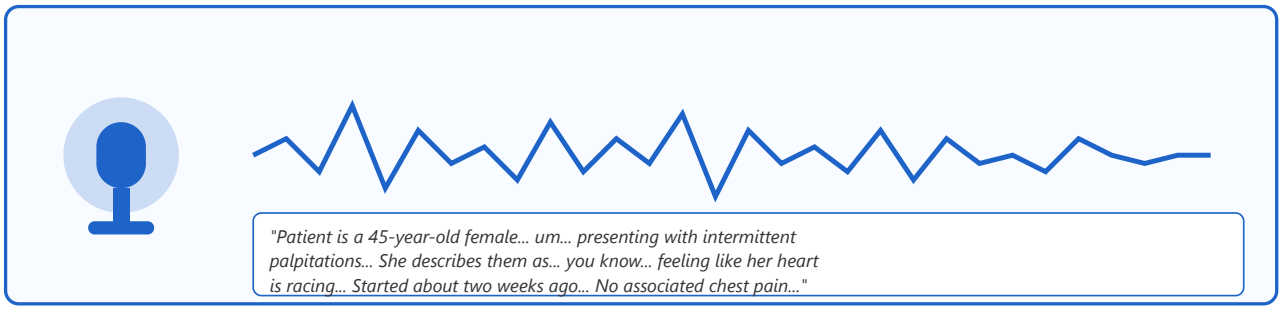
IMPRESSION:
1. Right lower lobe pneumonia
2. Small right pleural effusion

Electronically signed by:
Dr. Sarah Johnson, MD
Board Certified Radiologist



Multimodal Complexity: Radiology reports combine narrative text with images. NLP systems must extract findings like "3 cm opacity" and "right lower lobe" while also processing the actual radiographic images using computer vision techniques. The diagnostic impression requires understanding medical terminology and spatial relationships.

Example 3: Physician Voice Note



Speech Recognition Challenges: Voice recordings require automatic speech recognition (ASR) technology to transcribe spoken words, then NLP to clean up filler words ("um," "you know"), identify medical terms despite variations in pronunciation, and structure the content into meaningful clinical categories.

Example 4: Pathology Report with Microscopic Images

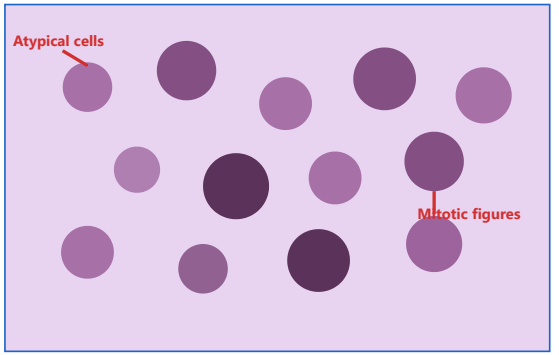
PATHOLOGY REPORT

SPECIMEN: Skin biopsy, right forearm

MICROSCOPIC DESCRIPTION:
Sections show epidermis with focal acanthosis and hyperkeratosis. The dermis reveals a proliferation of atypical melanocytes arranged in nests and sheets. Nuclear pleomorphism and increased mitotic figures are noted. Invasion into the reticular dermis is present with a Breslow depth of 1.2 mm.

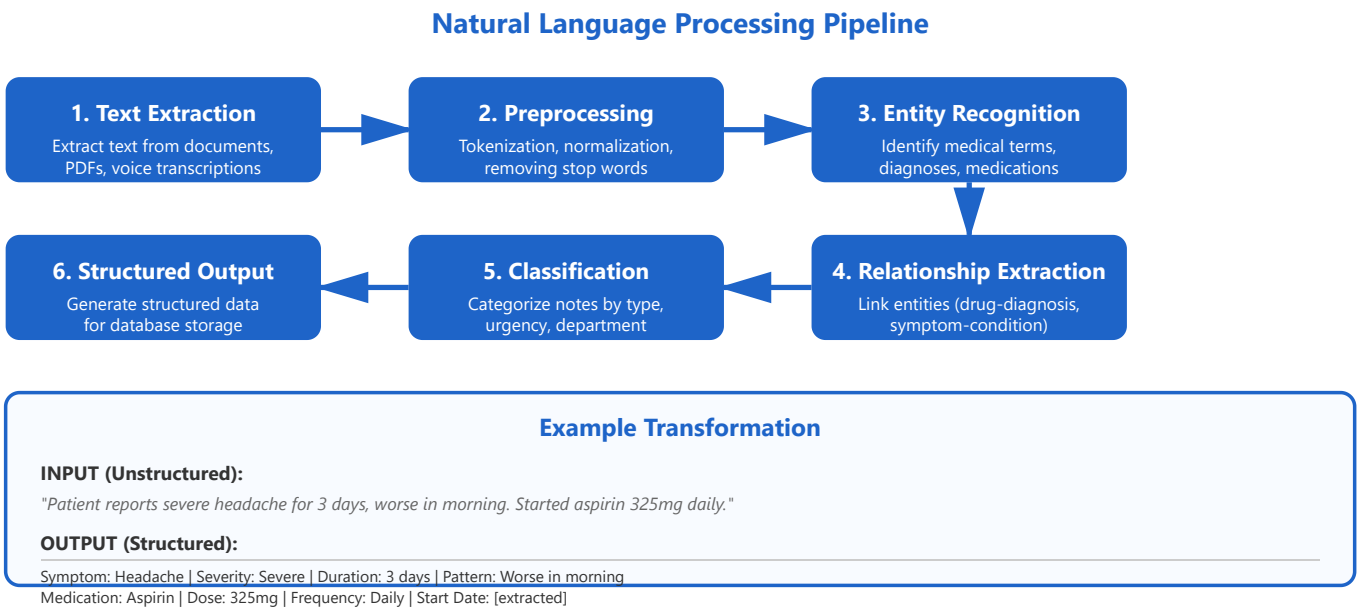
DIAGNOSIS:
Malignant melanoma, superficial spreading type
Clark Level III, Breslow 1.2mm

MICROSCOPIC IMAGE



Digital Pathology Analysis: Pathology reports combine highly technical narrative descriptions with microscopic images. AI systems must process both text (identifying terms like "Breslow depth" and "Clark Level") and images (detecting cellular abnormalities, counting mitotic figures) to support diagnostic accuracy and research.

▶ NLP Processing Pipeline



▶ Common Healthcare Unstructured Data Types

- **Clinical Notes:**Progress notes, discharge summaries, consultation reports, operative notes
- **Diagnostic Reports:**Radiology reports, pathology reports, cardiology interpretations
- **Medical Images:**X-rays, CT scans, MRIs, ultrasounds, pathology slides
- **Scanned Documents:**Historical paper records, consent forms, insurance documents
- **Communication:**Physician voice notes, patient messages, email correspondence

- **Social Data:**Patient-reported outcomes, survey responses, social determinants of health narratives
- **Multimedia:**Video recordings (telemedicine), audio recordings (patient interviews)

⚖️ Key Differences Summary

Aspect	Structured Data	Unstructured Data
Format	Predefined fields, tables, codes	Free text, images, audio, video
Storage	Relational databases (SQL)	Document stores, file systems, data lakes
Query Method	SQL queries, direct field access	Full-text search, NLP, AI analysis
Analysis Complexity	Simple aggregations, statistics	Requires NLP, machine learning, computer vision
Data Volume	~20% of healthcare data	~80% of healthcare data
Examples	Lab values, vital signs, ICD codes	Clinical notes, X-rays, pathology reports
Processing Speed	Fast (milliseconds)	Slow (seconds to minutes)
Standardization	High (standard codes, formats)	Low (variable expression, context-dependent)
Human Readability	Requires interpretation (codes)	Directly readable narratives

Aspect	Structured Data	Unstructured Data
Clinical Richness	Limited context	Rich clinical context and nuance

▶ Integration Challenges & Solutions

Challenge: Bridging Structured and Unstructured Data

Healthcare organizations must integrate both data types to gain comprehensive patient insights. For example, a diabetes care analysis requires structured glucose values AND unstructured clinical notes describing patient lifestyle, barriers to care, and treatment responses.

Solution: Unified Data Platforms

Modern healthcare data platforms combine traditional databases with document stores and NLP pipelines. Clinical data warehouses increasingly use hybrid architectures that store structured data in SQL databases while processing unstructured data through NLP engines, then linking results through patient identifiers and encounter IDs.