

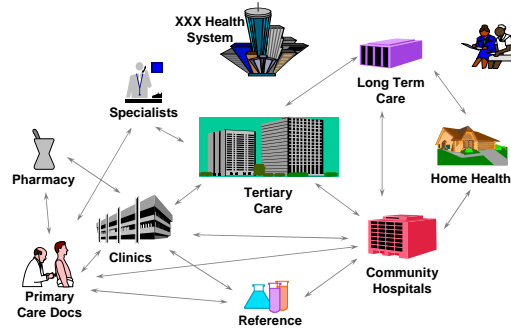
Health Enterprise Computing and Patient Identification

6.872/HST.950

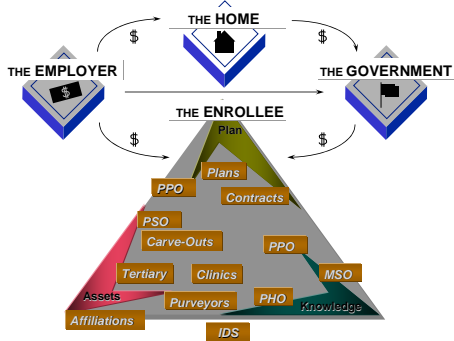
Peter Szolovits

(some content from David Margulies)

Integrated Delivery System...



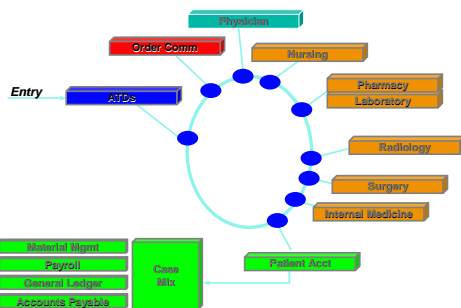
Complexity now within...



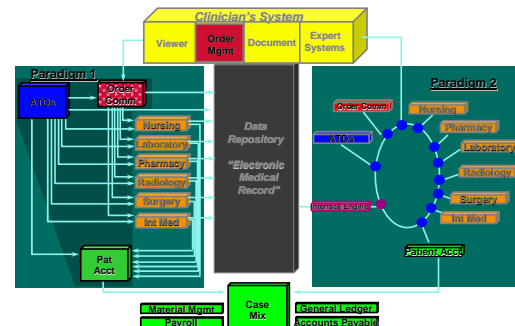
Context: Evolution of Clinical Computing...

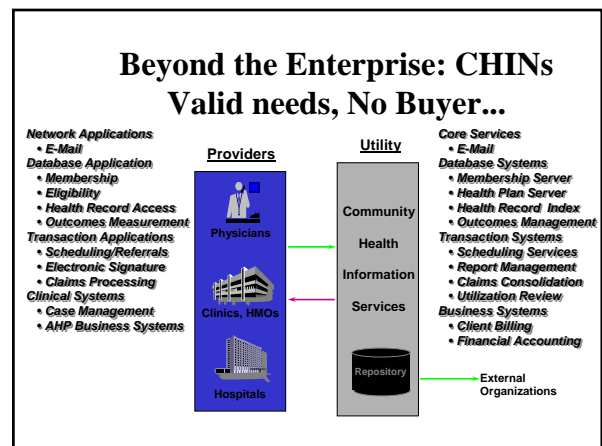
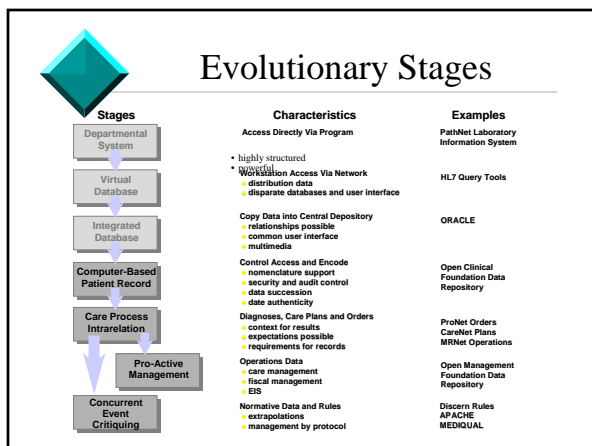
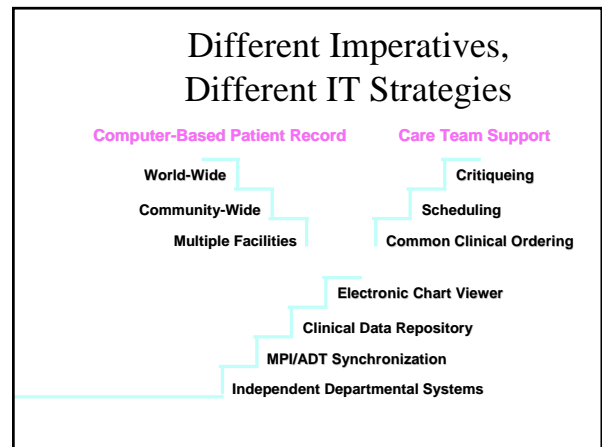
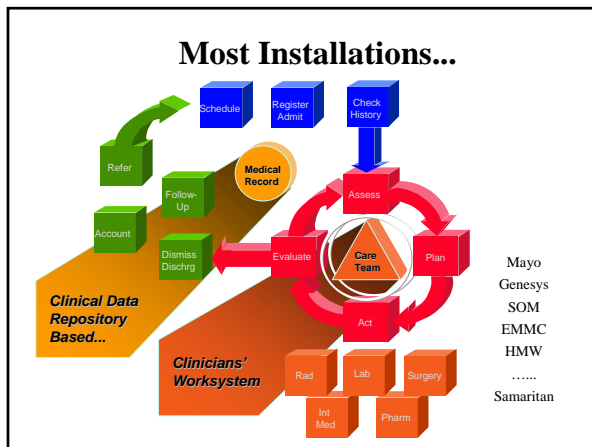
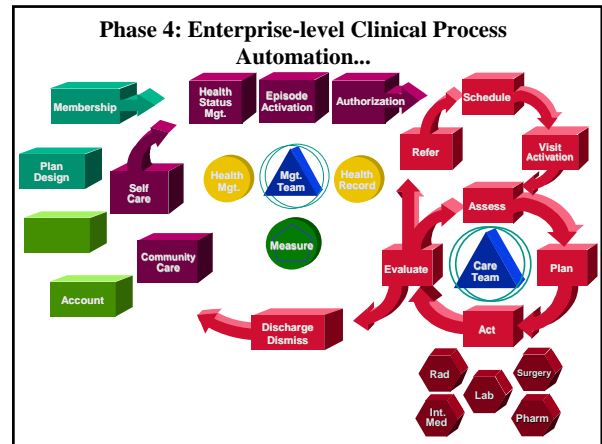
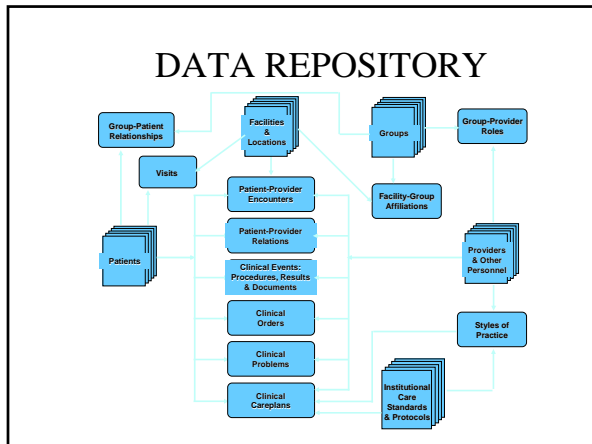
- Focus evolved from automating AR process in hospitals, to LAN-based SI of departmental systems, to enterprise clinical computing, toward region- and system-level process automation
- Changing healthcare organizational models and economic models ---> changing clinical computing priorities and designs

Phase 2: LAN-based SI -- Automate Data Distribution to MDs...



Phase 3: Data-based SI to Create an EMR...

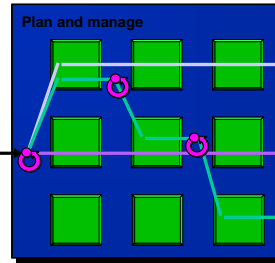




Community Health Information Needs

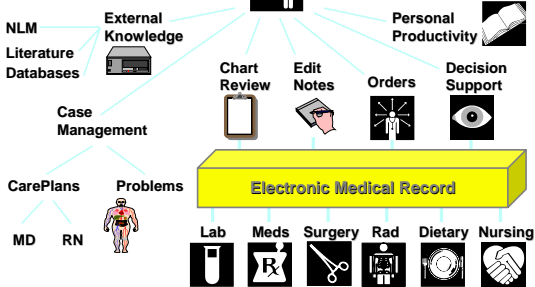
- Person Identification Server
- Membership Server
- Plan/Eligibility Server
- Health Record Server
- Resource Scheduling Server
- Care Protocol Server
- Commerce Server

Care Maps

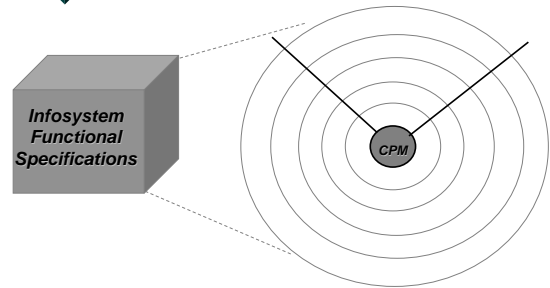


- Defined and tested clinical methods
- Specific paths through map individualized
- Multiple opportunities for triage and adaptation
 - by planned choice
 - by risk
 - by availability
 - by length of trajectory
 - by cost
 - by new information

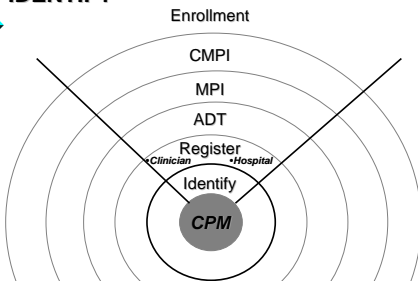
THE EYE OF THE CLINICIAN



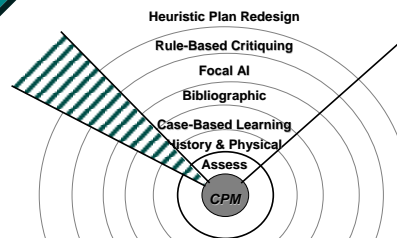
“EMERGING FROM THE COMMON PROCESS MODEL”

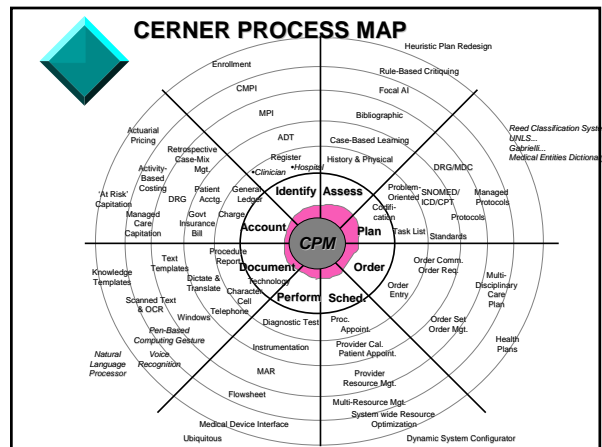
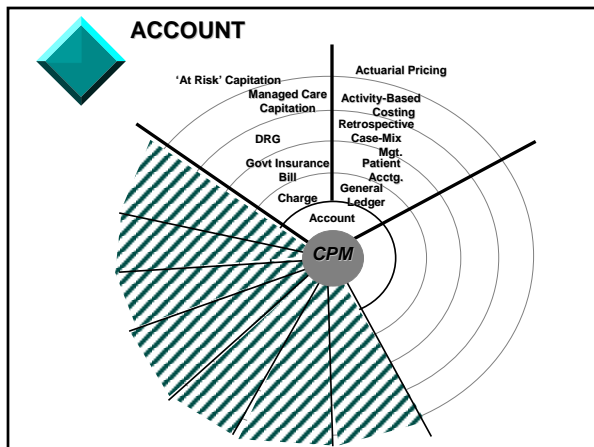
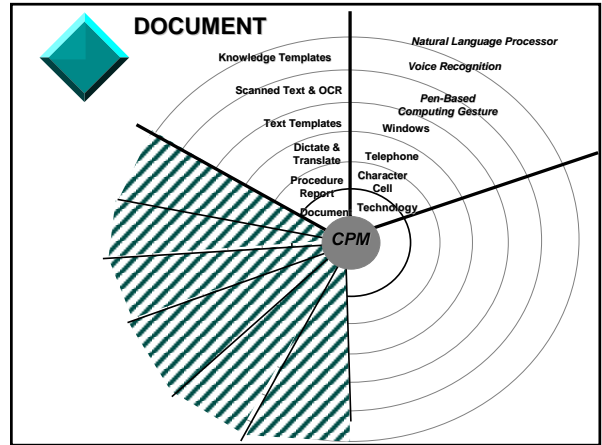
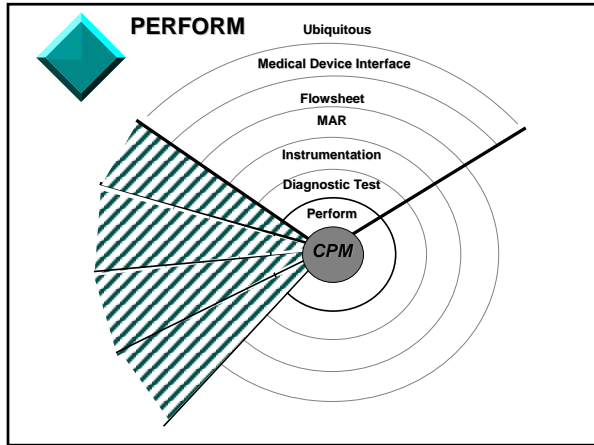
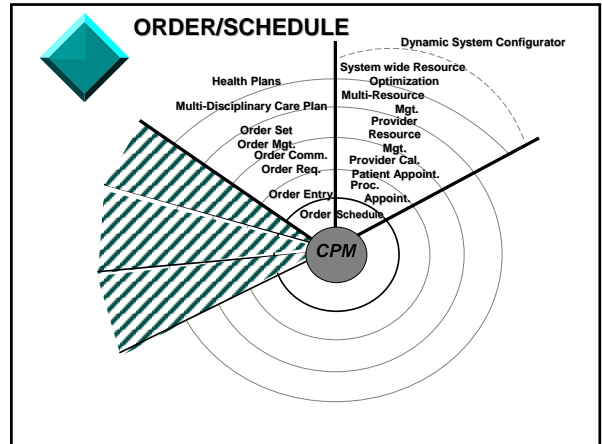
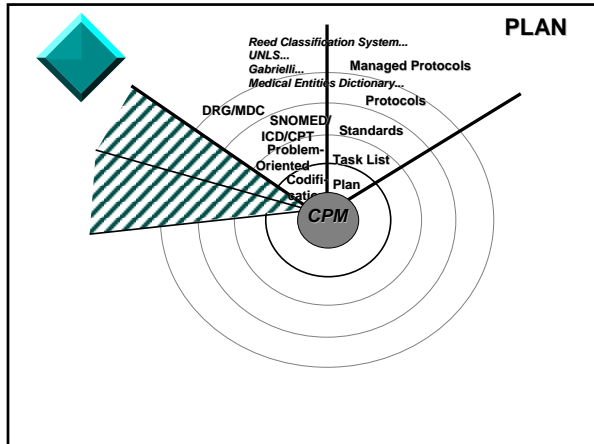


IDENTIFY



ASSESS





Patient Identification



Patient Identification

- Example: Partners Health Care EMPI (Enterprise Master Patient Index)
 - Lisa Adragna, J AHIMA 1998 Oct;69(9):46-8, 50, 52
- Probabilistic Models for Matching

Partners EMPI Goals

- (Context) Partners Healthcare is affiliation of Brigham and Women's, Mass. General, North-Shore Medical Center, Dana-Farber Cancer Institute, Spaulding Rehab, McLean, Partners Community Healthcare, ..., formed in 1994.
- Eliminate need to renumber existing systems by creating Partners-wide unique ID
- Foundation for Clinical Data Repository
- Minimally invasive to current operations
- Rapid installation

Partners EMPI Scope

- Medical record "crosswalk" across Partners
 - Local medical record numbers linked to Partners-wide ID
- Back-end, real-time, one-way interface
 - No modifications to existing systems
- Automated tools for identifying and resolving duplicates
 - No human intervention at "run time"

Partners EMPI Matching

- Three methods of lookup:
 - Last and First names
 - First name, Date of Birth and Gender
 - Social Security Number
- If no matches, create new EMPI number
- If exact match on all known features (above), then link to existing EMPI entry
- If multiple possible matches, create new entry and flag as potential duplicate; *manual intervention*

Partners EMPI Experience (1997)

- 1M records entered from BWH & MGH
- 170,000 pair of records identified as potential duplicates
- 80% of new registrations match a unique EMPI entry; ~1800/day
- Sites made 3308 "within-entity" merges
- EMPI Admin staff (3 FTE) resolved over 200,000 potential duplicates.

Models of Patient Matching

- *Ad hoc* methods are weak, and difficult to improve
- Motivational observation:
 - Two records that share the last name *Kowalczyk* are more likely to refer to the same individual than two that share the name *Smith*.

Computing posterior odds of a match

- $O(\text{match}/\text{obs}) = O(\text{match})L(\text{obs}/\text{match})$
- $O(\text{match}) = 1/N$
- $L(\text{obs}/\text{match}) = P(\text{obs}/\text{match})/P(\text{obs}/\sim\text{match})$
- $P(\text{Smith}, \text{Smith}|\text{match}) = P(\text{Smith})$
- $P(\text{Smith}, \text{Smith}|\sim\text{match}) = P(\text{Smith})^2$
- $L(\text{Smith}, \text{Smith}) = .01006/.01006^2 = 99.4$
- $L(\text{Kowalczyk}, \text{K}...) = 1/.00001 = 100,000$

Multiple evidence accumulates

- Smith has same Soundex code as Smyth
- L partially matches Liz
- Same street address number, street name
- Ave. vs. Blvd.
- Different phone numbers
- ...
- Assume conditional independence

Soundex

- Developed in 1800's to represent phonetic similarity of names
 - E.g., Smith \rightarrow S-530; Smyth \rightarrow S-530
- Rules:
 - 1 B, F, P, V
 - 2 C, G, J, K, Q, S, X, Z
 - 3 D, T
 - 4 L
 - 5 M, N
 - 6 R
 - Disregard the letters A, E, I, O, U, H, W, and Y.
 - Special rules for doubled consonants, prefixes, ...

Soundex examples

Szolovits	S-413
Beethoven	B-315
Smyth	S-530
Kowalczyk	K-422

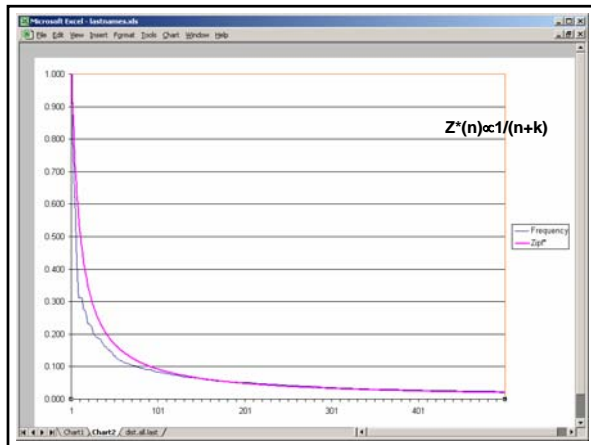
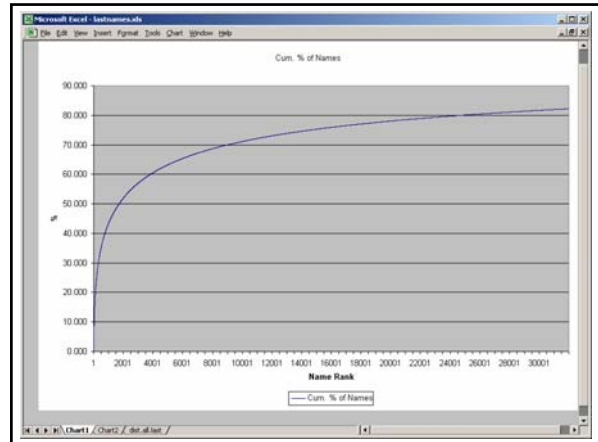
Typing/Transcription Error Model

- Consider probabilities of
 - Insertion
 - Deletion
 - Transposition
 - (Mutation) – varies with key layout
- Compute “distance” between mismatched features
- Same idea in **BLAST**

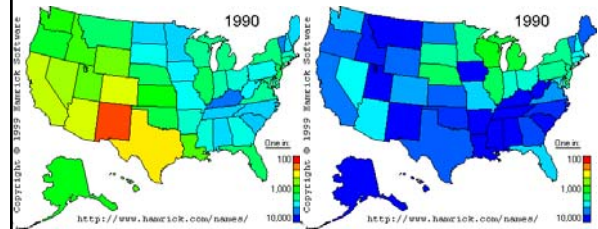
Frequency of U.S. Last Names

Rank	Name	Frequency	Rank
1	SMITH	1,000,000	1
2	JOHNSON	810,000	2
3	WILLIAMS	689,000	3
4	JONES	621,000	4
5	BROWN	521,000	5
6	DAVIS	480,000	6
7	MILLER	424,000	7
8	WILSON	339,000	8
9	MOORE	312,000	9
10	TAYLOR	311,000	10
11	ANDERSON	311,000	11
12	THOMAS	311,000	12
13	JACKSON	310,000	13
14	WHITE	279,000	14
15	HARRIS	275,000	15
16	MARTIN	273,000	16
17	THOMPSON	269,000	17
18	GARCIA	254,000	18
19	MARTINEZ	234,000	19
20	ROBINSON	223,000	20

--U.S. Census Bureau



Geographical Distribution of “Gonzales” vs. “Lewandowski”



A Theory for Record Linkage

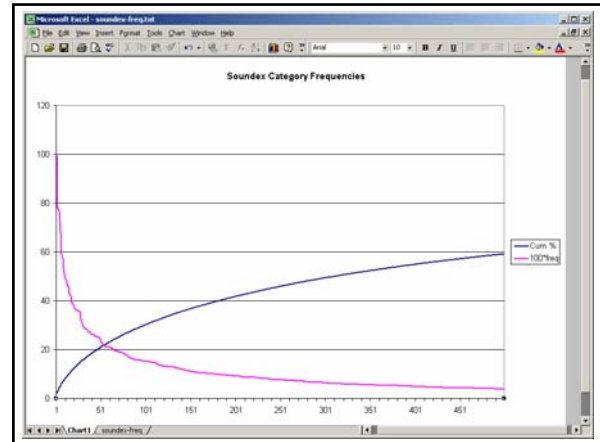
- Fellegi IP, Sunter AB. *J. Am. Stat. Assn.* **64**(328):1183-1210. 1969
- Estimate $L(obs/match)$ for all possible observations
 - Normally must make independence assumption to factor contributions of different aspects of *obs*
- Define *decision policy*
 - $P(A+|obs)$, $P(A?|obs)$, $P(A-|obs)$, $sum=1$
 - Based on thresholds on *L*'s
- Yields errors
 - $P(A+|~match)$ or $P(A-|match)$, minimize $P(A?)$

Blocking Sets

- Pair-by-pair comparison is expensive
 - E.g., 1M records at Partners $\rightarrow 10^{12}$ comparisons
- Therefore, compare only smaller subgroups
 - E.g., assume first letter of name is never misspelled, as Soundex does
 - E.g., assume Soundex pronunciation code is correct, even if spelling is not
- Increased chance of missing matches, much faster

4234 Soundex Classes /5616 possible

Smith, Schmidt, Schmitt, Sneed, Schmid, Smoot, Snead, Shumate, Smyth, Sandy, Sand, Santo, Santoyo, ...	S-530	1.118%
Johnson, Jenkins, Jensen, Jamison, Jameson, Jansen, Johansen, Janssen, ...	J-525	0.998%
Brown, Bryan, Barron, Braun, Byrne, Bruno, Browne, Baron, Barney, Burney...	B-650	0.779%
Jones, James, Johns, Jung, Janes, Jaynes, Jonas, Jenks, Jaimes, Jinks, Jahnke, Johannes, Janik, Janis, Jiang, ...	J-520	0.775%



Themes

- Growing levels of integration in health care
- Demand for integration of health IT systems raises challenging problems
- MPI is fundamental for record keeping
- Matching of records critical
 - Good theory, mixed practice
- Flip side: privacy concerns