Big Data Analytics for Healthcare

Jimeng Sun

Healthcare Analytics Department

IBM TJ Watson Research Center

Chandan K. Reddy

Department of Computer Science

Wayne State University

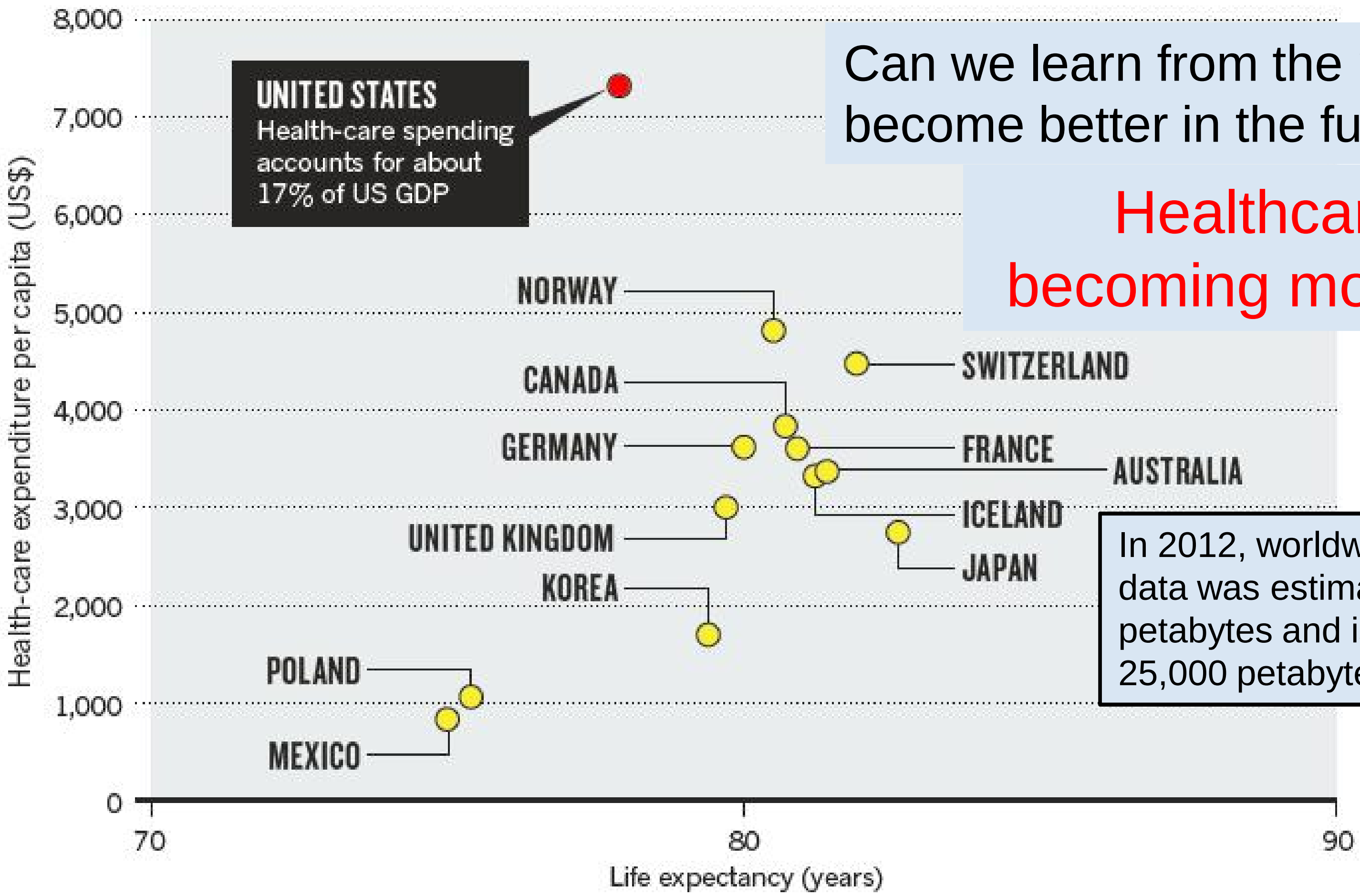
Tutorial presentation at the SIAM International Conference on Data Mining, Austin, TX, 2013.

The updated tutorial slides are available at http://dmkd.cs.wayne.edu/TUTORIAL/Healthcare/

Motivation

MONEY WELL SPENT?

The United States has not seen an increase in life expectancy to match its huge outlay on health care.



UNITED STATES
Health-care spending accounts for about 17% of US GDP

Can we learn from the past to become better in the future ??

Healthcare Data is becoming more complex !!

In 2012, worldwide digital healthcare data was estimated to be equal to 500 petabytes and is expected to reach 25,000 petabytes in 2020.

Hersh, W., Jacko, J. A., Greenes, R., Tan, J., Janies, D., Embi, P. J., & Payne, P. R. (2011). Health-care hit or miss? *Nature*, *470*(7334), 327.

Organization of this Tutorial

- Introduction
- Motivating Examples
- Sources and Techniques for Big Data in Healthcare
  - Structured EHR Data
  - Unstructured Clinical Notes
  - Medical Imaging Data
  - Genetic Data
  - Other Data (Epidemiology & Behavioral)
- Final Thoughts and Conclusion

INTRODUCTION

Definition of Big Data

- A collection of large and complex data sets which are difficult to process using common database management tools or traditional data processing applications.

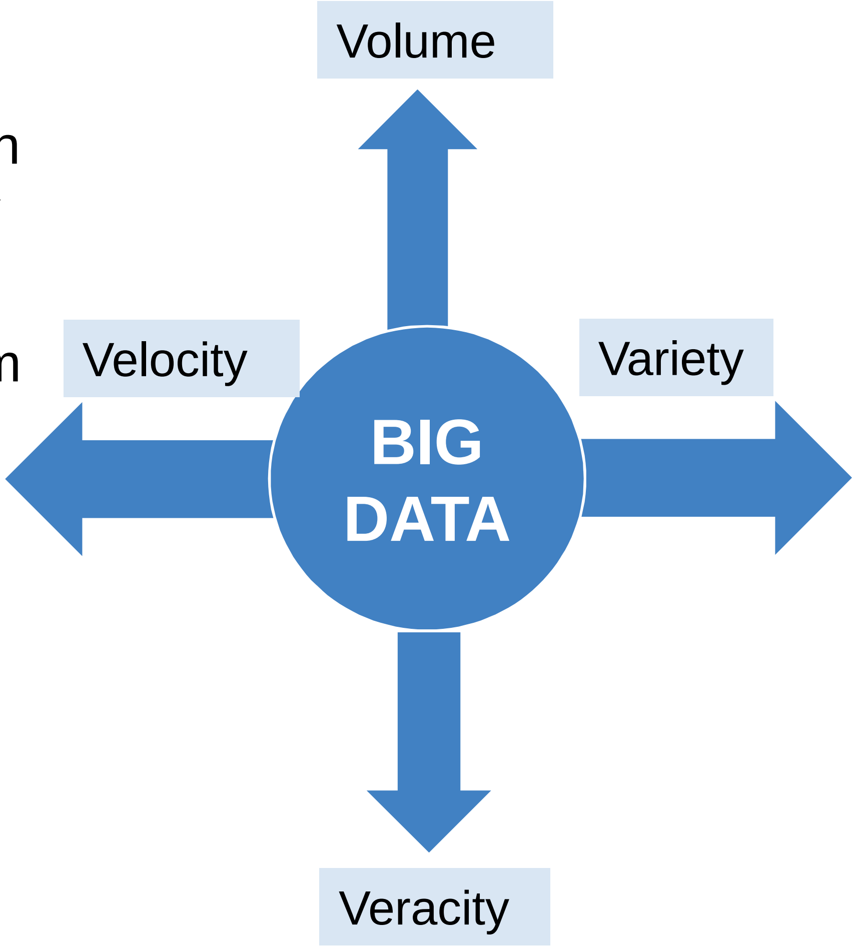
- “Big data refers to the tools, processes and procedures allowing an organization to create, manipulate, and manage very large data sets and storage facilities”

– according to zdnet.com

Big data is not just about size.

- Finds insights from complex, noisy, heterogeneous, longitudinal, and voluminous data.
- It aims to answer questions that were previously unanswered.

The challenges include capturing, storing, searching, sharing & analyzing.



The four dimensions (V’s) of Big Data

Reasons for Growing Complexity/Abundance of Healthcare Data

- Standard medical practice is moving from relatively ad-hoc and subjective decision making to evidence-based healthcare.
- More incentives to professionals/hospitals to use EHR technology.

Additional Data Sources

- Development of new technologies such as capturing devices, sensors, and mobile applications.
- Collection of genomic information became cheaper.
- Patient social communications in digital forms are increasing.
- More medical knowledge/discoveries are being accumulated.

Big Data Challenges in Healthcare

- Inferring knowledge from complex heterogeneous patient sources. Leveraging the patient/data correlations in longitudinal records.

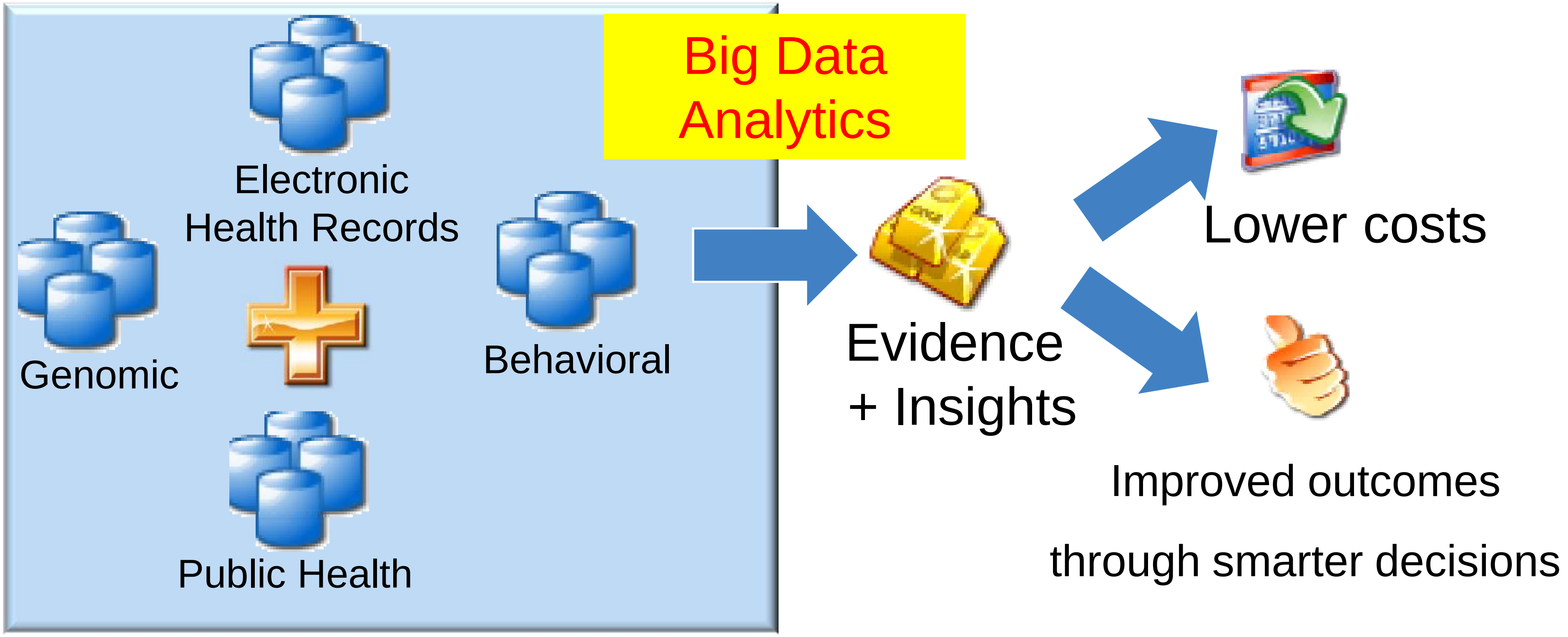
- Understanding unstructured clinical notes in the right context.

- Efficiently handling large volumes of medical imaging data and extracting potentially useful information and biomarkers.

- Analyzing genomic data is a computationally intensive task and combining with standard clinical data adds additional layers of complexity.

- Capturing the patient’s behavioral data through several sensors; their various social interactions and communications.

Overall Goals of Big Data Analytics in Healthcare



- Take advantage of the massive amounts of data and provide right intervention to the right patient at the right time.
- Personalized care to the patient.
- Potentially benefit all the components of a healthcare system i.e., provider, payer, patient, and management.

Purpose of this Tutorial

Two-fold objectives:

- Introduce the data mining researchers to the sources available and the possible challenges and techniques associated with using big data in healthcare domain.
- Introduce Healthcare analysts and practitioners to the advancements in the computing field to effectively handle and make inferences from voluminous and heterogeneous healthcare data.

The ultimate goal is to bridge data mining and medical informatics communities to foster interdisciplinary works between the two communities.

PS: Due to the broad nature of the topic, the primary emphasis will be on introducing healthcare data repositories, challenges, and concepts to data scientists. Not much focus will be on describing the details of any particular techniques and/or solutions.

Disclaimers

- Being a recent and growing topic, there might be several other resources that might not be covered here.
- Presentation here is more biased towards the data scientists' perspective and may be less towards the healthcare management or healthcare provider's perspective.
- Some of the website links provided might become obsolete in the future. This tutorial is prepared in early 2013.
- Since this topic contains a wide varieties of problems, there might be some aspects of healthcare that might not be covered in the tutorial.

MOTIVATING EXAMPLES

EXAMPLE 1: Heritage Health Prize

Improve Healthcare,
Win $3,000,000.

Identify patients who will be admitted to a hospital within the next year using historical claims data.

http://www.heritagehealthprize.com

- Over $30 billion was spent on unnecessary hospital admissions.

Goals:

- Identify patients at high-risk and ensure they get the treatment they need.
- Develop algorithms to predict the number of days a patient will spend in a hospital in the next year.

Outcomes:

- Health care providers can develop new strategies to care for patients before its too late → reduces the number of unnecessary hospitalizations.
- Improving the health of patients while decreasing the costs of care.
- Winning solutions use a combination of several predictive models.

EXAMPLE 2: Penalties for Poor Care - 30-Day Readmissions

- Hospitalizations account for more than 30% of the 2 trillion annual cost of healthcare in the United States. Around 20% of all hospital admissions occur within 30 days of a previous discharge.
  - not only expensive but are also potentially harmful, and most importantly, they are often preventable.

- Medicare penalizes hospitals that have high rates of readmissions among patients with heart failure, heart attack, and pneumonia.

- Identifying patients at risk of readmission can guide efficient resource utilization and can potentially save millions of healthcare dollars each year.

- Effectively making predictions from such complex hospitalization data will require the development of novel advanced analytical models.

EXAMPE 3: White House unveils BRAIN Initiative

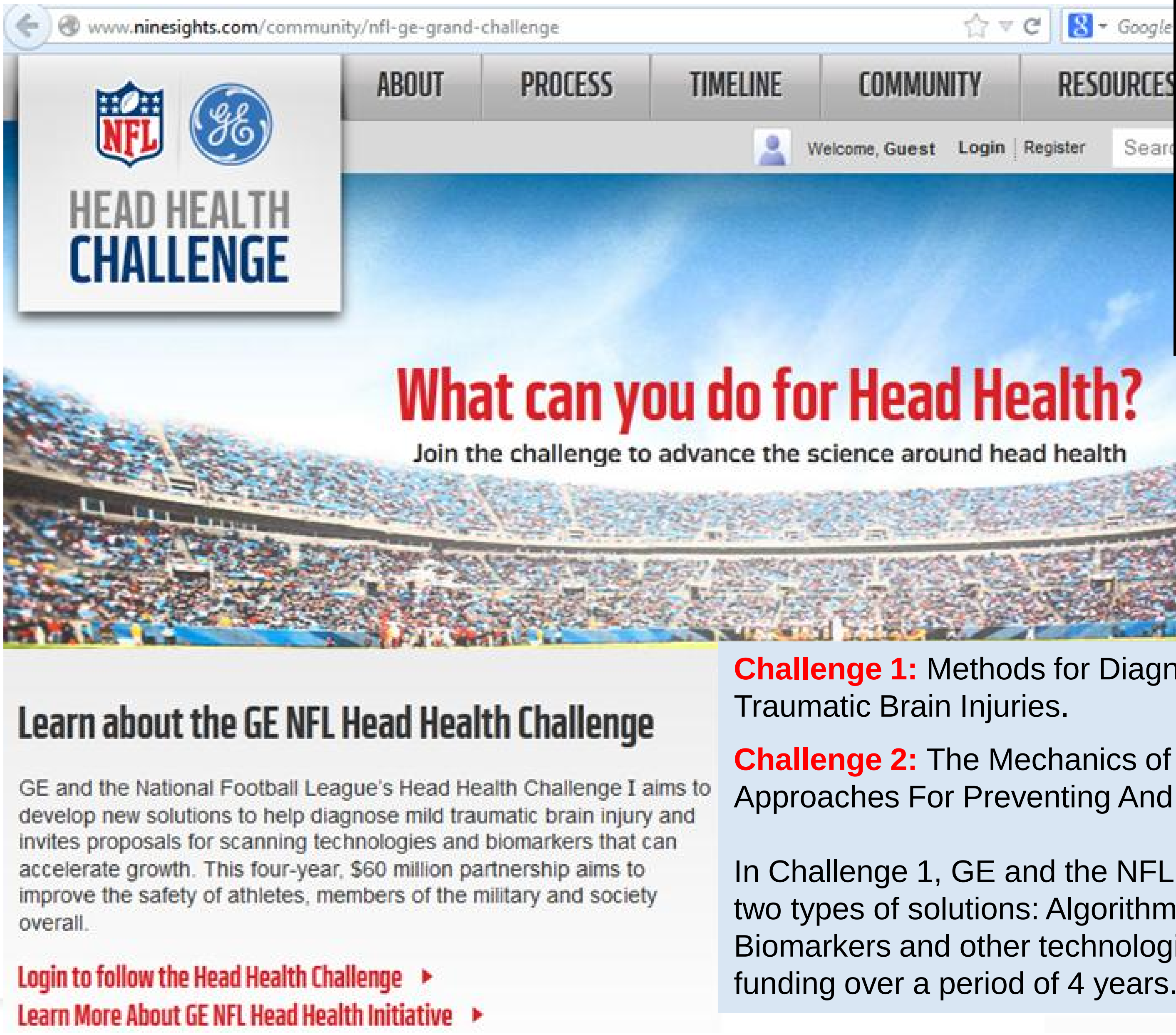
- The US President unveiled a new bold $100 million research initiative designed to revolutionize our understanding of the human brain. BRAIN (Brain Research through Advancing Innovative Neurotechnologies) Initiative.
- Find new ways to treat, cure, and even prevent brain disorders, such as Alzheimer’s disease, epilepsy, and traumatic brain injury.
- *“Every dollar we invested to map the human genome returned $140 to our economy... Today, our scientists are mapping the human brain to unlock the answers to Alzheimer’s.*”

  -- President Barack Obama, 2013 State of the Union.

- “advances in "Big Data" that are necessary to analyze the huge amounts of information that will be generated; and increased understanding of how thoughts, emotions, actions and memories are represented in the brain.” : NSF
- Joint effort by NSF, NIH, DARPA, and other private partners.

http://www.whitehouse.gov/infographics/brain-initiative

EXAMPLE 4: GE Head Health Challenge



What can you do for Head Health?

Join the challenge to advance the science around head health

Learn about the GE NFL Head Health Challenge

GE and the National Football League's Head Health Challenge I aims to develop new solutions to help diagnose mild traumatic brain injury and invites proposals for scanning technologies and biomarkers that can accelerate growth. This four-year, $60 million partnership aims to improve the safety of athletes, members of the military and society overall.

Login to follow the Head Health Challenge ▸

Learn More About GE NFL Head Health Initiative ▸

Challenge 1: Methods for Diagnosis and Prognosis of Mild Traumatic Brain Injuries.

Challenge 2: The Mechanics of Injury: Innovative Approaches For Preventing And Identifying Brain Injuries.

In Challenge 1, GE and the NFL will award up to $10M for two types of solutions: Algorithms and Analytical Tools, and Biomarkers and other technologies. A total of $60M in funding over a period of 4 years.

Healthcare Continuum

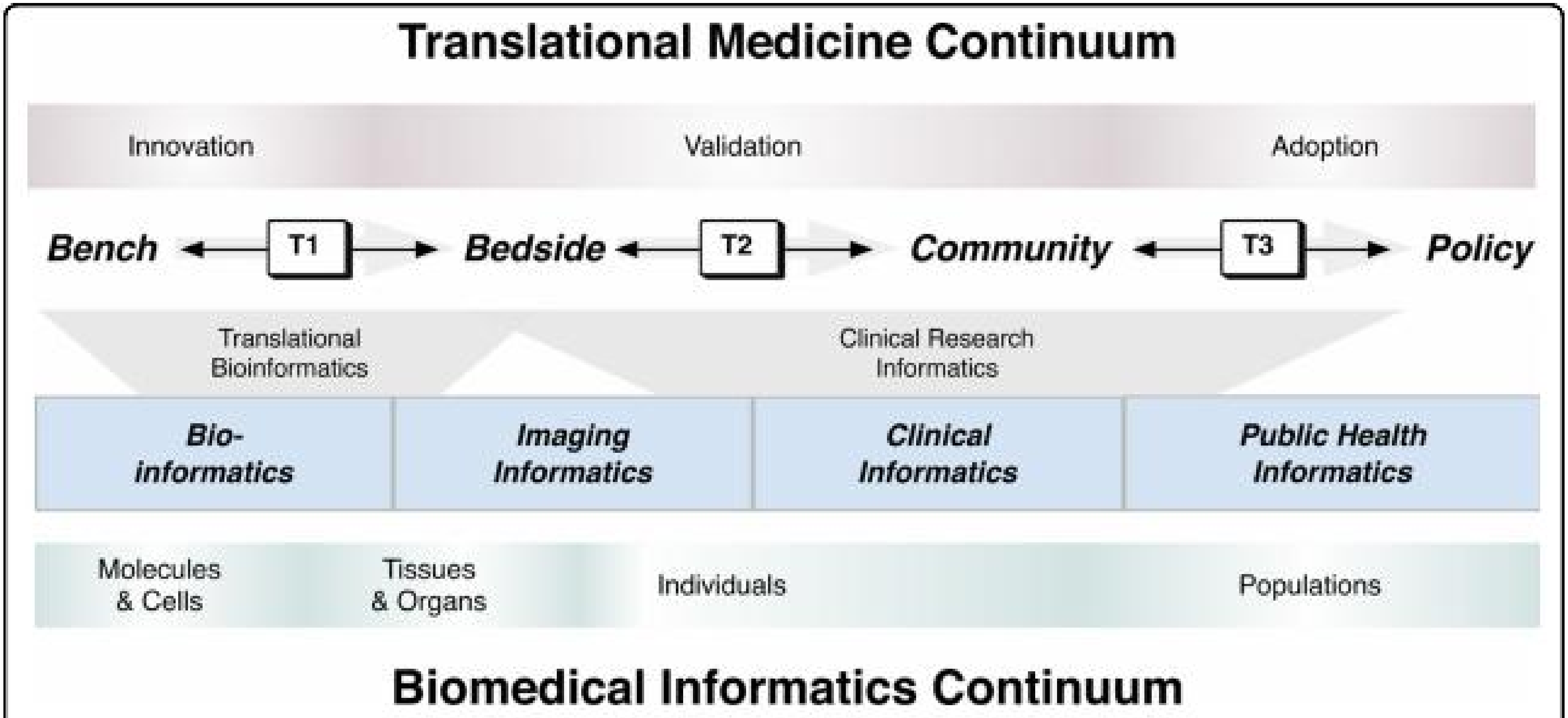
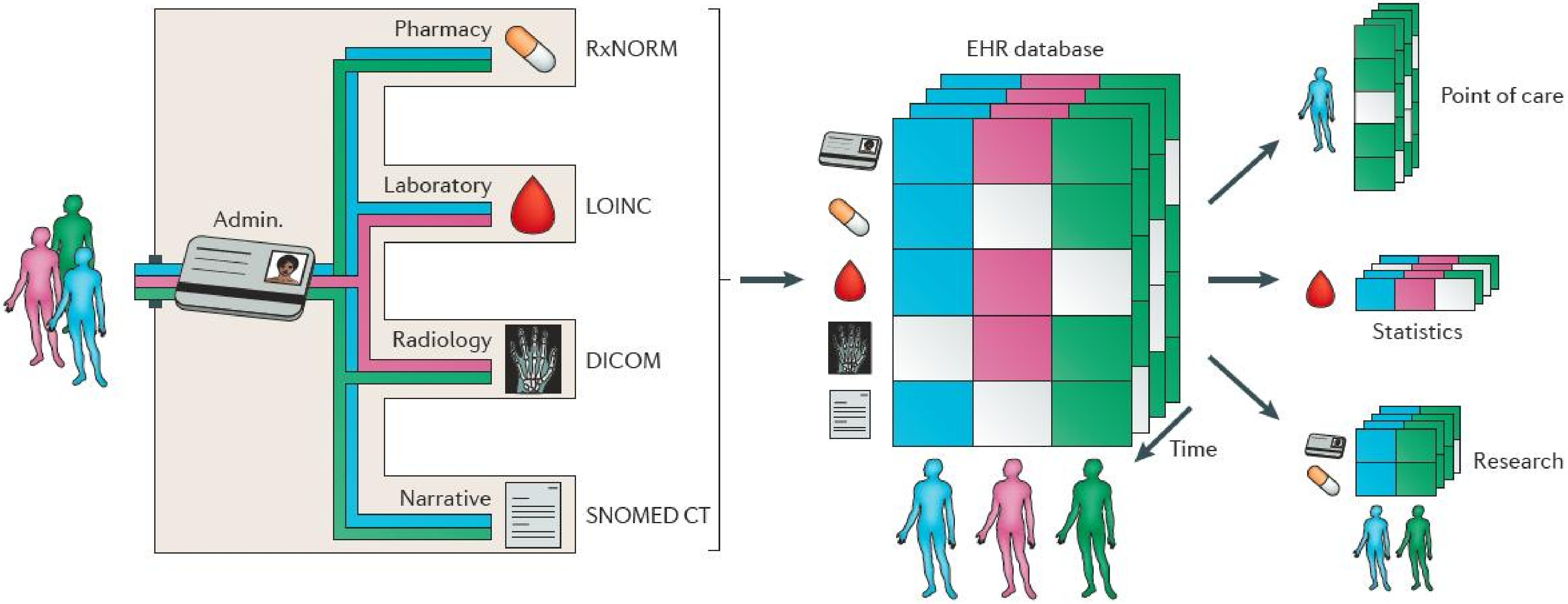


Figure 1 The synergistic relationship across the biomedical informatics and translational medicine continua. Major areas of translational medicine (along the top; innovation, validation, and adoption) are depicted relative to core focus areas of biomedical informatics (along the bottom; molecules and cells, tissues and organs, individuals, and populations). The crossing of translational barriers (T1, T2, and T3) can be enabled using translational bioinformatics and clinical research informatics approaches, which are comprised of methodologies from across the sub-disciplines of biomedical informatics (e.g., bioinformatics, imaging informatics, clinical informatics, and public health informatics).

Sarkar, Indra Neil. "Biomedical informatics and translational medicine." *Journal of Translational Medicine* 8.1 (2010): 22.

Data Collection and Analysis



Effectively integrating and efficiently analyzing various forms of healthcare data over a period of time can answer many of the impending healthcare problems.

Jensen, Peter B., Lars J. Jensen, and Søren Brunak. "Mining electronic health records: towards better research applications and clinical care." *Nature Reviews Genetics* (2012).

Organization of this Tutorial

- **Introduction**
- **Motivating Examples**
- **Sources and Techniques for Big Data in Healthcare**
  - **Structured EHR Data**
  - **Unstructured Clinical Notes**
  - **Medical Imaging Data**
  - **Genetic Data**
  - **Other Data (Epidemiology & Behavioral)**
- **Final Thoughts and Conclusion**

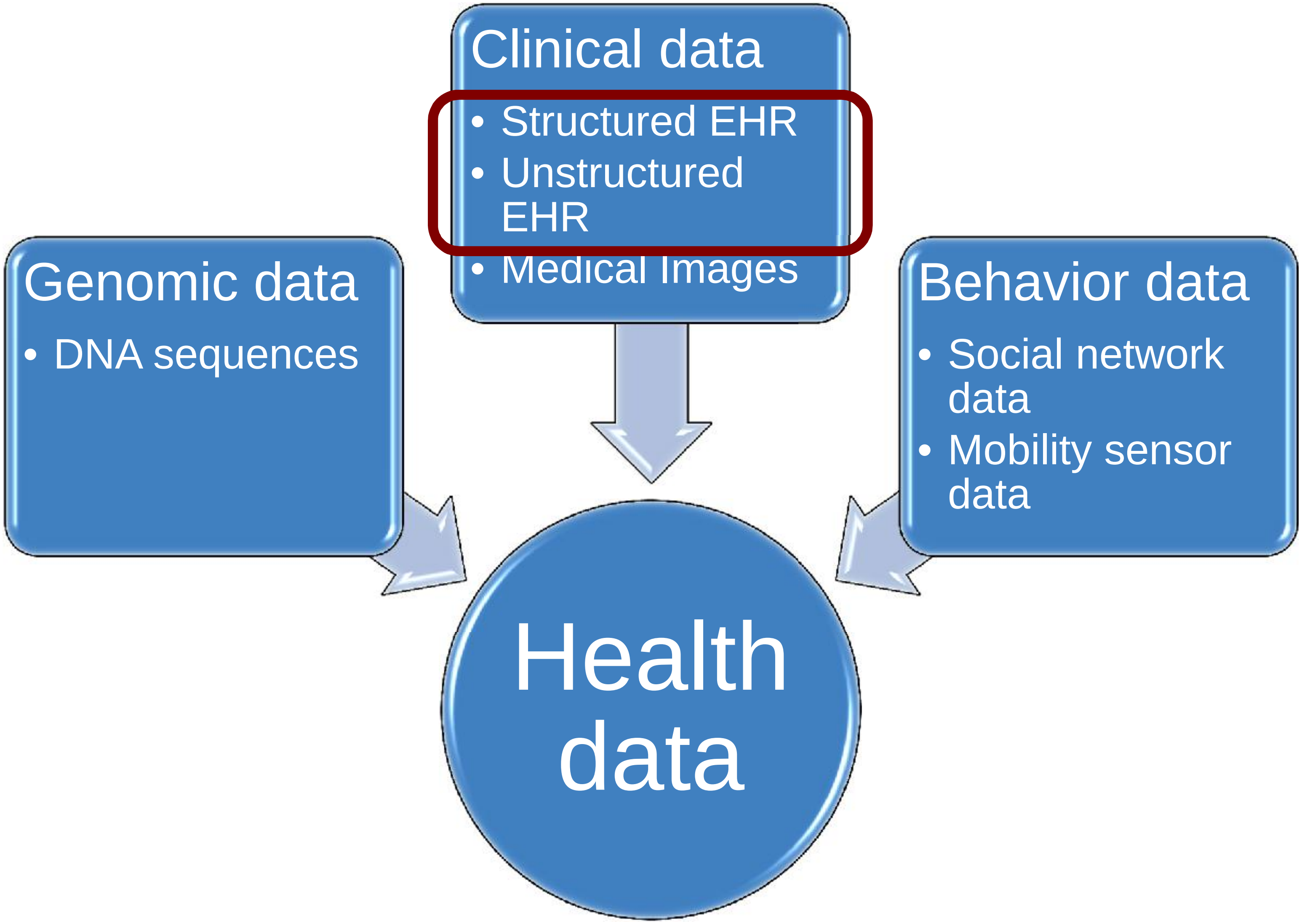
SOURCES AND TECHNIQUES FOR BIG DATA IN HEALTHCARE

Outline

- Electronic Health Records (EHR) data
- Healthcare Analytic Platform
- Resources

ELECTRONIC HEALTH RECORDS (EHR) DATA

Data



Clinical data

- Structured EHR
- Unstructured EHR
- Medical Images

Genomic data

- DNA sequences

Behavior data

- Social network data
- Mobility sensor data

Health data

Billing data - ICD codes

- ICD stands for International Classification of Diseases
- ICD is a hierarchical terminology of diseases, signs, symptoms, and procedure codes maintained by the World Health Organization (WHO)
- In US, most people use ICD-9, and the rest of world use ICD-10
- Pros: Universally available
- Cons: medium recall and medium precision for characterizing patients

- (250) Diabetes mellitus
    - (250.0) Diabetes mellitus without mention of complication
    - (250.1) Diabetes with ketoacidosis
    - (250.2) Diabetes with hyperosmolarity
    - (250.3) Diabetes with other coma
    - (250.4) Diabetes with renal manifestations
    - (250.5) Diabetes with ophthalmic manifestations
    - (250.6) Diabetes with neurological manifestations
    - (250.7) Diabetes with peripheral circulatory disorders
    - (250.8) Diabetes with other specified manifestations
    - (250.9) Diabetes with unspecified complication

Billing data – CPT codes

- CPT stands for Current Procedural Terminology created by the American Medical Association
- CPT is used for billing purposes for clinical services
- Pros: High precision
- Cons: Low recall

Codes for Evaluation and Management: 99201-99499
(99201 - 99215) office/other outpatient services
(99217 - 99220) hospital observation services
(99221 - 99239) hospital inpatient services
(99241 - 99255) consultations
(99281 - 99288) emergency dept services
(99291 - 99292) critical care services
…

Lab results

- The standard code for lab is Logical Observation Identifiers Names and Codes (LOINC®)
- Challenges for lab
  - Many lab systems still use local dictionaries to encode labs
  - Diverse numeric scales on different labs
    - Often need to map to normal, low or high ranges in order to be useful for analytics
  - Missing data
    - not all patients have all labs
    - The order of a lab test can be predictive, for example, BNP indicates high likelihood of heart failure

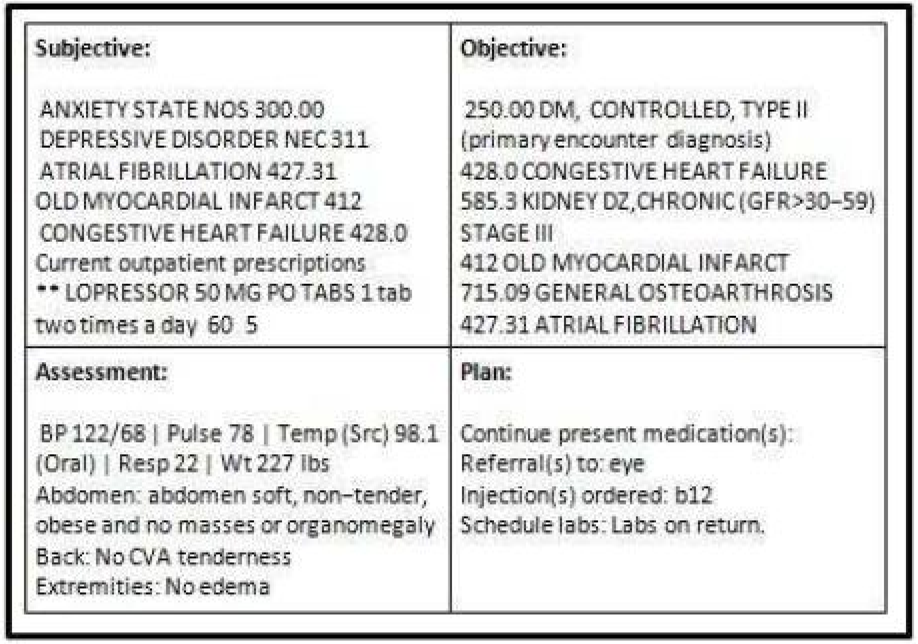
Time	Lab	Value
1996-03-15 12:50:00.0	CO2	29.0
1996-03-15 12:50:00.0	BUN	16.0
1996-03-15 12:50:00.0	HDL-C	37.0
1996-03-15 12:50:00.0	K	4.5
1996-03-15 12:50:00.0	Cl	102.0
1996-03-15 12:50:00.0	Gluc	86.0

Medication

- Standard code is National Drug Code (NDC) by Food and Drug Administration (FDA), which gives a unique identifier for each drug
  - Not used universally by EHR systems
  - Too specific, drugs with the same ingredients but different brands have different NDC
- RxNorm: a normalized naming system for generic and branded drugs by National Library of Medicine
- Medication data can vary in EHR systems
  - can be in both structured or unstructured forms
- Availability and completeness of medication data vary
  - Inpatient medication data are complete, but outpatient medication data are not
  - Medication usually only store prescriptions but we are not sure whether patients actually filled those prescriptions

Clinical notes

- Clinical notes contain rich and diverse source of information
- Challenges for handling clinical notes
  - Ungrammatical, short phrases
  - Abbreviations
  - Misspellings
  - Semi-structured information
    - Copy-paste from other structure source
      - Lab results, vital signs
    - Structured template:
      - SOAP notes: Subjective, Objective, Assessment, Plan



Subjective: ANXIETY STATE NOS 300.00 DEPRESSIVE DISORDER NEC 311 ATRIAL FIBRILLATION 427.31 OLD MYOCARDIAL INFARCT 412 CONGESTIVE HEART FAILURE 428.0 Current outpatient prescriptions ** LOPRESSOR 50 MG PO TABS 1 tab two times a day 60 5	**Objective:** 250.00 DM, CONTROLLED, TYPE II (primary encounter diagnosis) 428.0 CONGESTIVE HEART FAILURE 585.3 KIDNEY DZ,CHRONIC (GFR>30-59) STAGE III 412 OLD MYOCARDIAL INFARCT 715.09 GENERAL OSTEOARTHROSIS 427.31 ATRIAL FIBRILLATION
Assessment: BP 122/68 | Pulse 78 | Temp (Src) 98.1 (Oral) | Resp 22 | Wt 227 lbs Abdomen: abdomen soft, non-tender, obese and no masses or organomegaly Back: No CVA tenderness Extremities: No edema	**Plan:** Continue present medication(s): Referral(s) to: eye Injection(s) ordered: b12 Schedule labs: Labs on return.

Summary of common EHR data

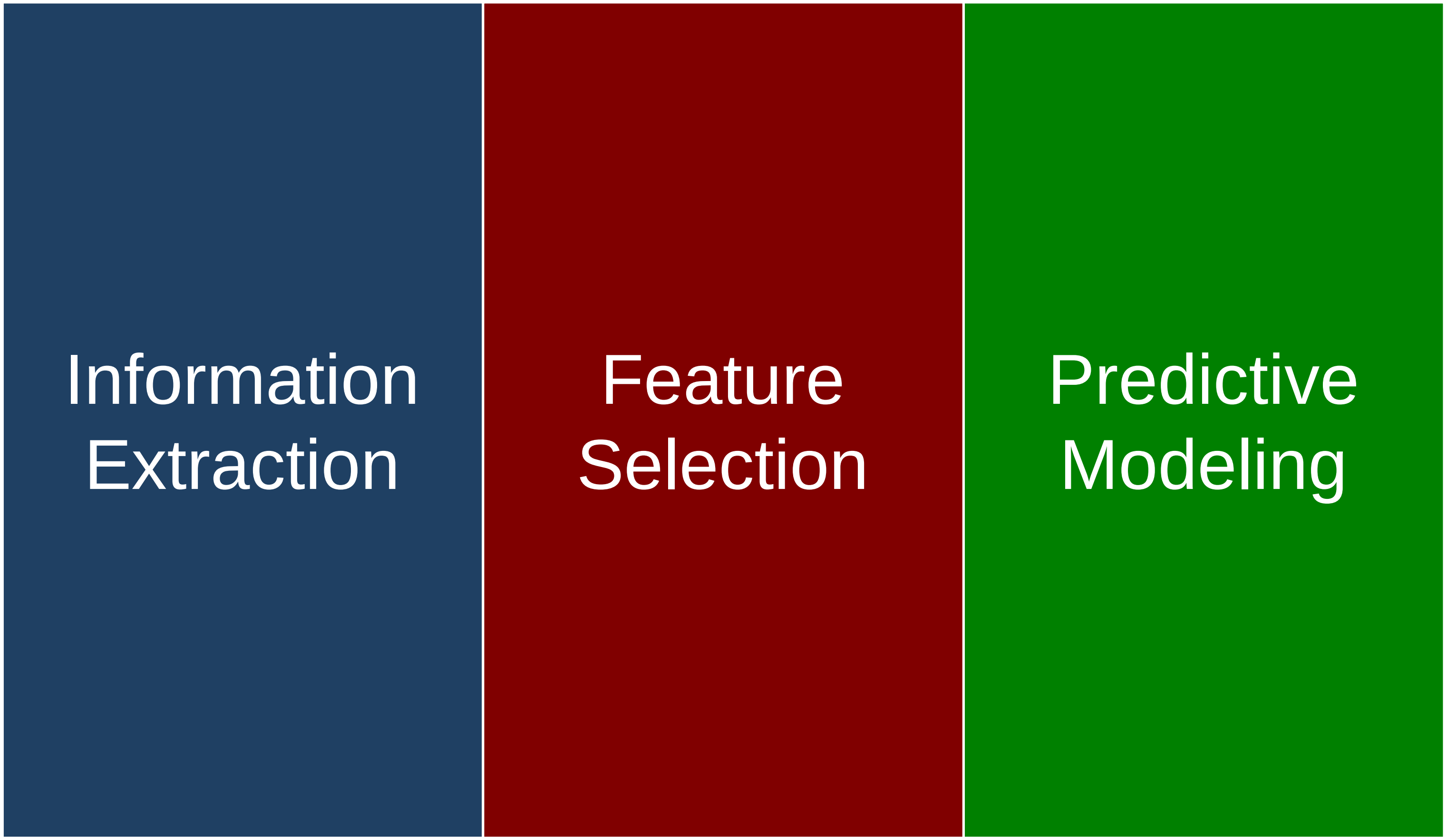
	ICD	CPT	Lab	Medication	Clinical notes
Availability	High	High	High	Medium	Medium
Recall	Medium	Poor	Medium	Inpatient: High Outpatient: Variable	Medium
Precision	Medium	High	High	Inpatient: High Outpatient: Variable	Medium high
Format	Structured	Structured	Mostly structured	Structured and unstructured	Unstructured
Pros	Easy to work with, a good approximation of disease status	Easy to work with, high precision	High data validity	High data validity	More details about doctors' thoughts
Cons	Disease code often used for screening, therefore disease might not be there	Missing data	Data normalization and ranges	Prescribed not necessary taken	Difficult to process

Joshua C. Denny Chapter 13: Mining Electronic Health Records in the Genomics Era. PLoS Comput Biol. 2012 December; 8(12):

Analytic Platform

Large-scale Healthcare Analytic Platform

Analytic Platform

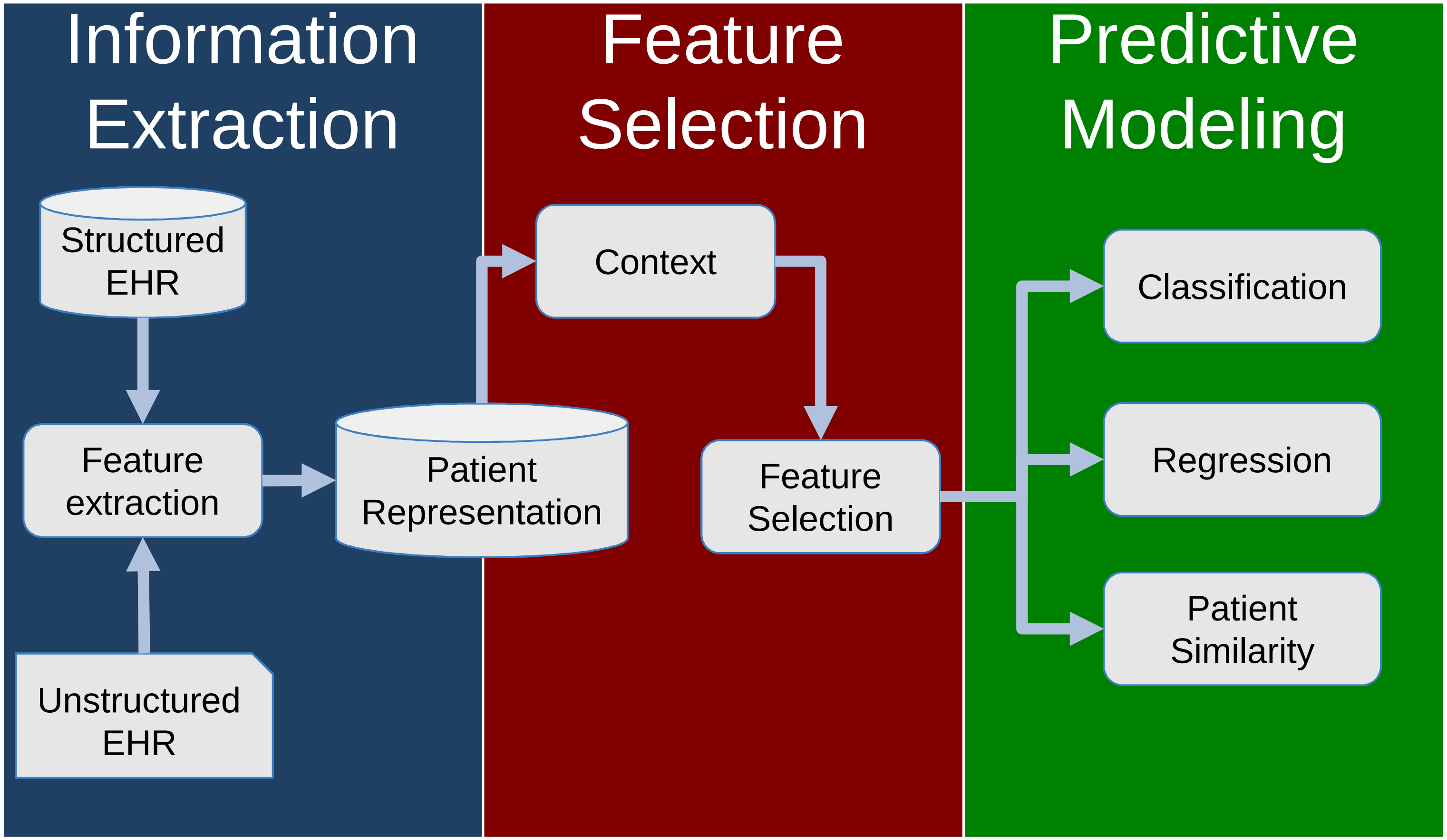


Information Extraction

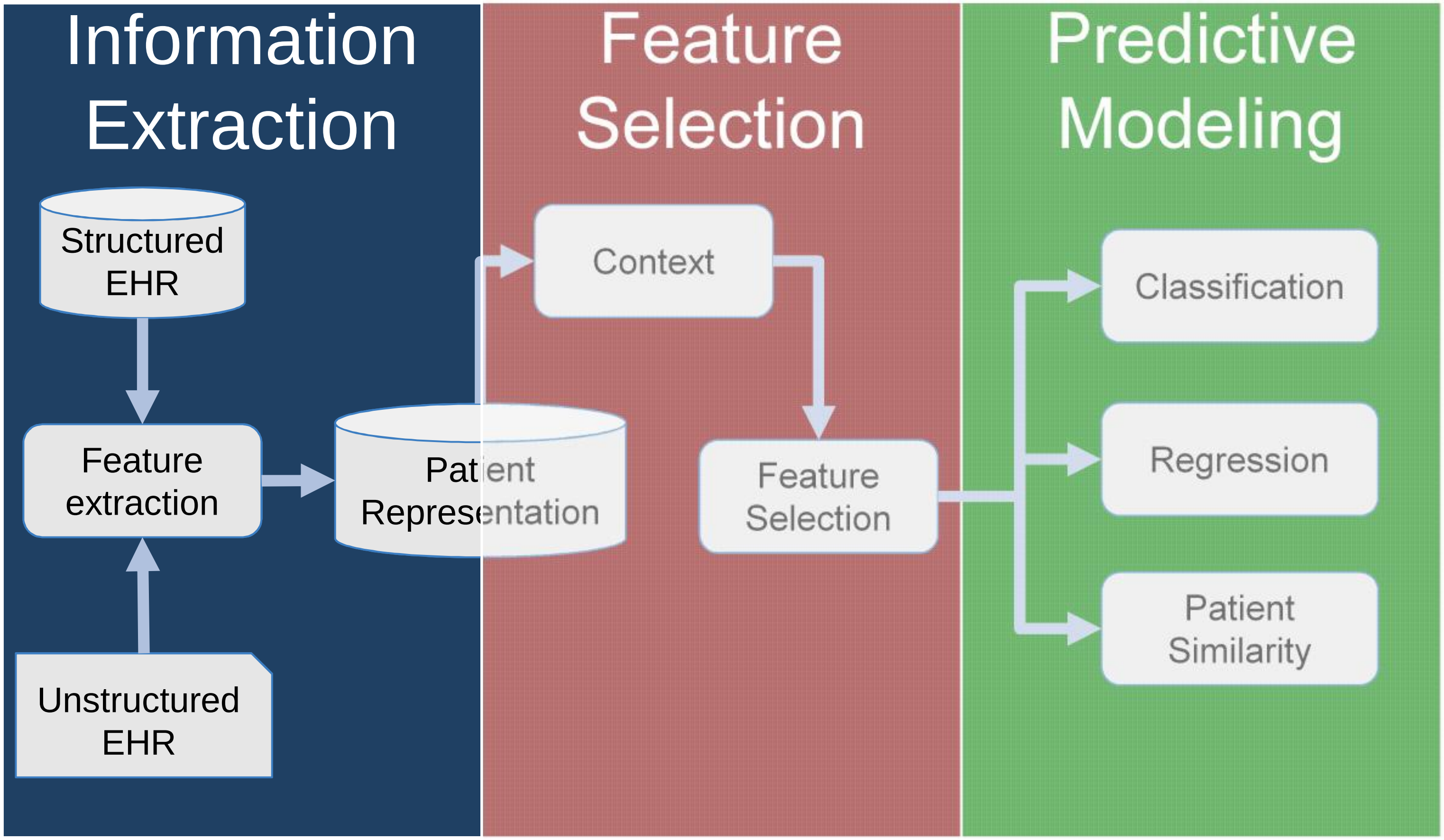
Feature Selection

Predictive Modeling

Analytic Platform



Analytic Platform



CLINICAL TEXT MINING

Text Mining in Healthcare

- Text mining
  - – Information Extraction
    - • Name Entity Recognition
  - – Information Retrieval
- Clinical text vs. Biomedical text
  - – Biomedical text: medical literatures (well-written medical text)
  - – Clinical text is written by clinicians in the clinical settings

- Meystre et al. Extracting Information from Textual Documents in the Electronic Health Record: A Review of Recent Research. IMIA 2008
- Zweigenbaum et al. Frontiers of biomedical text mining: current progress, BRIEFINGS IN BIOINFORMATICS. VOL 8. NO 5. 358-375
- Cohen and Hersh, A survey of current work in biomedical text mining. BRIEFINGS IN BIOINFORMATICS. VOL 6. NO 1. 57–71.

Auto-Coding: Extracting Codes from Clinical Text

- Problem
  - Automatically assign diagnosis codes to clinical text
- Significance
  - **The cost is approximately $25 billion per year in the US**
- Available Data
  - Medical NLP Challenges from 2007
    - Subsections from radiology reports: clinical history and impression
- Potential Evaluation Metric:
  - F-measure = 2P*R/(P+R), where P is precision, and R is recall.
- Example References

- Aronson et al. From indexing the biomedical literature to coding clinical text: experience with MTI and machine learning approaches. BioNLP 2007
- Crammer et al. Automatic Code Assignment to Medical Text. BioNLP 2007
- Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. JAMIA. 2004:392-402
- Pakhomov SV, Buntrock JD, Chute CG. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. JAMIA 2006:516-25.

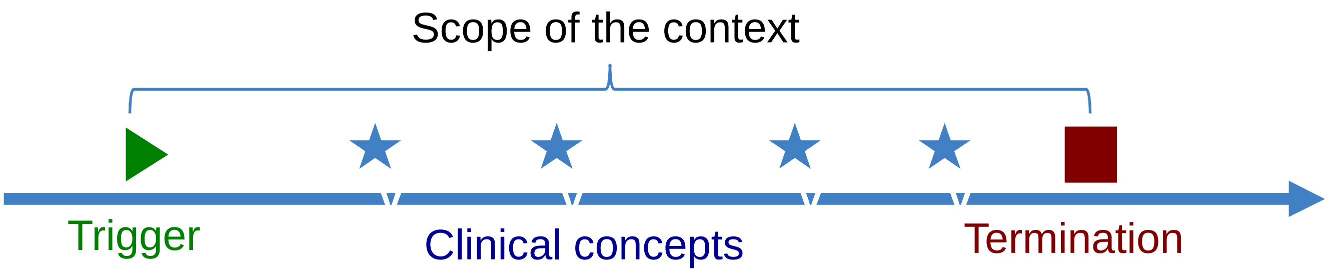
Context Analysis - Negation

- Negation: e.g., ...denies chest pain…
  - NegExpander [1] achieves 93% precision on mammographic reports
  - NegEx [2] uses regular expression and achieves 94.5% specificity and 77.8% sensitivity
  - NegFinder [3] uses UMLS and regular expression, and achieves 97.7 specificity and 95.3% sensitivity when analyzing surgical notes and discharge summaries
  - A hybrid approach [4] uses regular expression and grammatical parsing and achieves 92.6% sensitivity and 99.8% specificity

1. Aronow DB, Fangfang F, Croft WB. Ad hoc classification of radiology reports. JAMIA 1999:393-411
2. Chapman et al. A simple algorithm for identifying negated findings and diseases in discharge summaries. JBI 2001:301-10.
3. Mutalik PG, et al. Use of general-purpose negation detection to augment concept indexing of medical documents: a quantitative study using the UMLS. JAMIA 2001:598-609.
4. Huang Y, Lowe HJ. A novel hybrid approach to automated negation detection in clinical radiology reports. JAMIA 2007

Context Analysis - Temporality

- Temporality: e.g., …fracture of the tibia 2 years ago
  - – TimeText [1] can detect temporal relations with 93.2% recall and 96.9% precision on 14 discharge summaries
  - – Context [2] is an extension of NegEx, which identifies
    - negations (negated, affirmed),
    - temporality (historical, recent, hypothetical)
    - experiencer (patient, other)



1. Zhou et al. The Evaluation of a Temporal Reasoning System in Processing Clinical Discharge Summaries. JAMIA 2007.
2. Chapman W, Chu D, Dowling JN. ConText: An Algorithm for Identifying Contextual Features from Clinical Text. BioNLP 2007

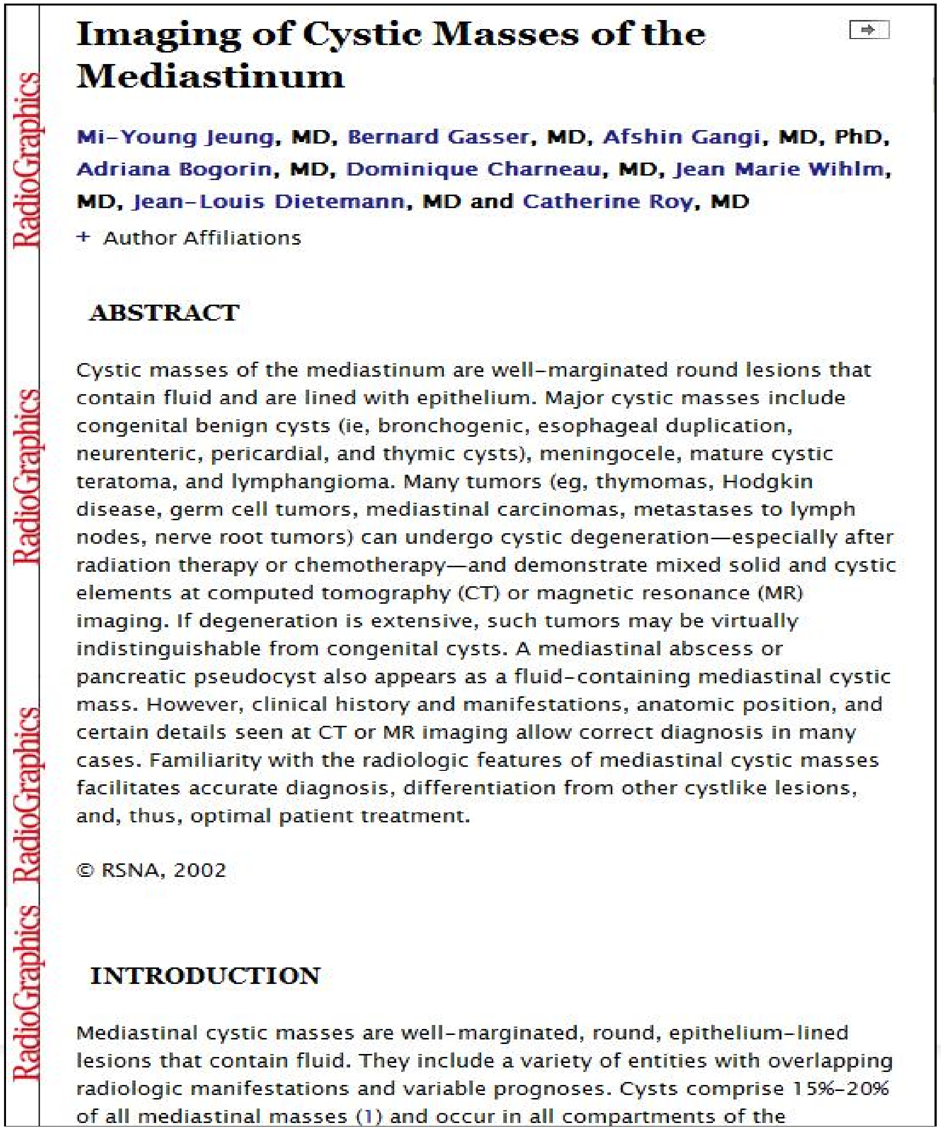
CASE 1: CASE BASED RETRIEVAL

Sondhi P, Sun J, Zhai C, Sorrentino R, Kohn MS. Leveraging medical thesauri and physician feedback for improving medical literature retrieval for case queries. JAMIA. 2012

Input: Case Query

Patient with smoking habit and weight loss. The frontal and lateral chest X rays show a mass in the posterior segment of the right upper lobe as well as a right hilar enlargement and obliteration of the right paratracheal stripe. On the chest CT the contours of the mass are lobulated with heterogeneous enhancement.Enlarged mediastinal and hilar lymph nodes are present.

Goal: Find Relevant Research Articles to a Query



Imaging of Cystic Masses of the Mediastinum

Mi-Young Jeung, MD, Bernard Gasser, MD, Afshin Gangi, MD, PhD, Adriana Bogorin, MD, Dominique Charneau, MD, Jean Marie Wihlm, MD, Jean-Louis Dietemann, MD and Catherine Roy, MD

+ Author Affiliations

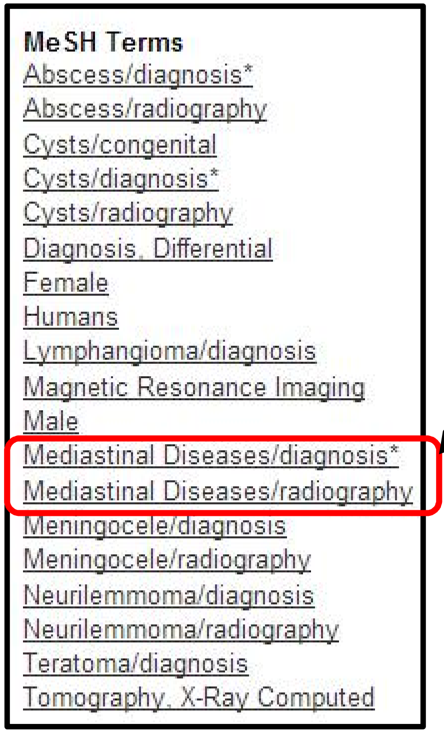
ABSTRACT

Cystic masses of the mediastinum are well-marginated round lesions that contain fluid and are lined with epithelium. Major cystic masses include congenital benign cysts (ie, bronchogenic, esophageal duplication, neurenteric, pericardial, and thymic cysts), meningocele, mature cystic teratoma, and lymphangioma. Many tumors (eg, thymomas, Hodgkin disease, germ cell tumors, mediastinal carcinomas, metastases to lymph nodes, nerve root tumors) can undergo cystic degeneration—especially after radiation therapy or chemotherapy—and demonstrate mixed solid and cystic elements at computed tomography (CT) or magnetic resonance (MR) imaging. If degeneration is extensive, such tumors may be virtually indistinguishable from congenital cysts. A mediastinal abscess or pancreatic pseudocyst also appears as a fluid-containing mediastinal cystic mass. However, clinical history and manifestations, anatomic position, and certain details seen at CT or MR imaging allow correct diagnosis in many cases. Familiarity with the radiologic features of mediastinal cystic masses facilitates accurate diagnosis, differentiation from other cystlike lesions, and, thus, optimal patient treatment.

INTRODUCTION

Mediastinal cystic masses are well-marginated, round, epithelium-lined lesions that contain fluid. They include a variety of entities with overlapping radiologic manifestations and variable prognoses. Cysts comprise 15%-20% of all mediastinal masses (1) and occur in all compartments of the

Additional Related Information



MeSH Terms

- Abscess/diagnosis*
- Abscess/radiography
- Cysts/congenital
- Cysts/diagnosis*
- Cysts/radiography
- Diagnosis, Differential
- Female
- Humans
- Lymphangioma/diagnosis
- Magnetic Resonance Imaging
- Male
- Mediastinal Diseases/diagnosis*
- Mediastinal Diseases/radiography
- Meningocele/diagnosis
- Meningocele/radiography
- Neurilemmoma/diagnosis
- Neurilemmoma/radiography
- Teratoma/diagnosis
- Tomography, X-Ray Computed

Disease MeSH

Challenge 1: Query Weighing

Patient with **smoking** habit and **weight loss**. The frontal and lateral **chest X rays** show a **mass** in the posterior segment of the **right upper lobe** as well as a right **hilar enlargement** and **obliteration of the right paratracheal stripe**. On the chest CT the contours of the mass are lobulated with heterogeneous enhancement. Enlarged mediastinal and hilar lymph nodes are present.

- Queries are long
- Not all words useful
- IDF does not reflect importance
- Semantics decide weight

Method: Semantic Query Weighing

Included UMLS semantic types

Disease or syndrome, Body part organ or organ component, Sign or symptom, Finding, Acquired abnormality, Congenital abnormality, Mental or behavioral dysfunction, Neoplasm, Pharmacologic substance, Individual Behavior

- Identify important UMLS Semantic Types based on their definition
- Assign higher weights to query words under these types

Challenge 2: Vocabulary Gap

Patient with **smoking** habit and **weight loss**. The frontal and lateral **chest X rays** show a **mass** in the posterior segment of the **right upper lobe** as well as a right **hilar enlargement** and **obliteration of the right paratracheal stripe**. On the chest CT the contours of the mass are lobulated with heterogeneous enhancement.Enlarged mediastinal and hilar lymph nodes are present.

- Matching variants
  - “*x ray*”, “*x-rays*”, “*x rays*”
- Matching synonyms
  - “*CT*” or “*x rays*”
- Knowledge gap

Method: Additional Query Keywords

Female patient, 25 years old, with fatigue and a swallowing disorder (dysphagia worsening during a meal). The frontal chest X-ray shows opacity with clear contours in contact with the right heart border. Right hilar structures are visible through the mass. The lateral X-ray confirms the presence of a mass in the anterior mediastinum. On CT images, the mass has a relatively homogeneous tissue density.

Additional keywords: Thymoma, Lymphoma, Dysphagia, Esophageal obstruction, Myasthenia gravis, Fatiguability, Ptosis

- Asked physicians to provide additional keywords
- Adding them with low weight helps
- Any potential diagnosis keywords help greatly
- Gives us insights into better query formulation

Challenge 3: Pseudo-Feedback

- General vocabulary gap solution:
  - Apply Pseudo-Relevance Feedback

- What if very few of the top N are relevant?

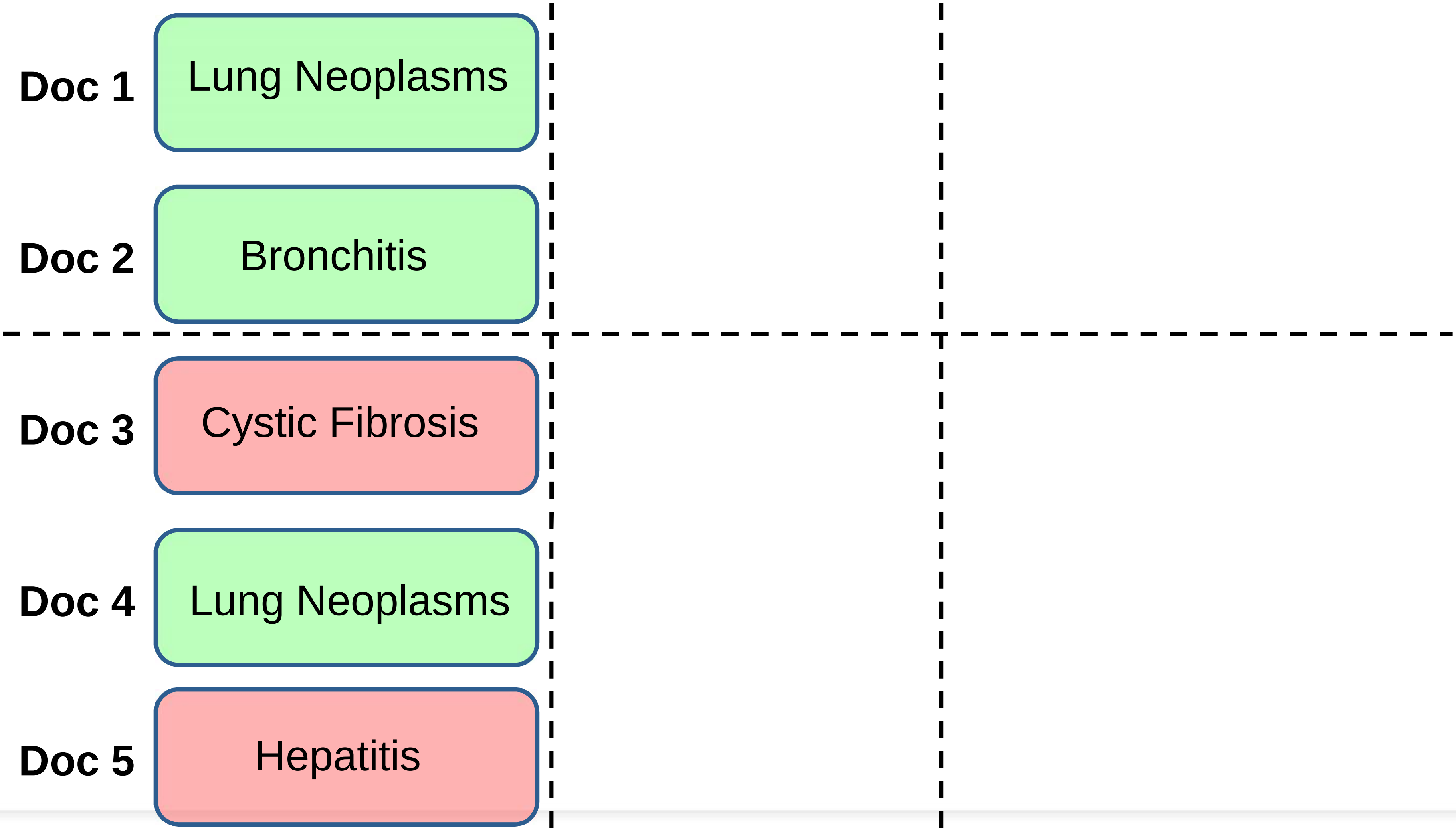
- No idea which keywords to pick up

Method: Medical Subject Heading (MeSH) Feedback

- Any case related query usually relates only to handful of conditions

- How to guess the condition of the query?
  - Select MeSH terms from top N=10 ranked documents
  - Select MeSH terms covering most query keywords
  - Use them for feedback

Method: MeSH Feedback



Doc 1 Lung Neoplasms

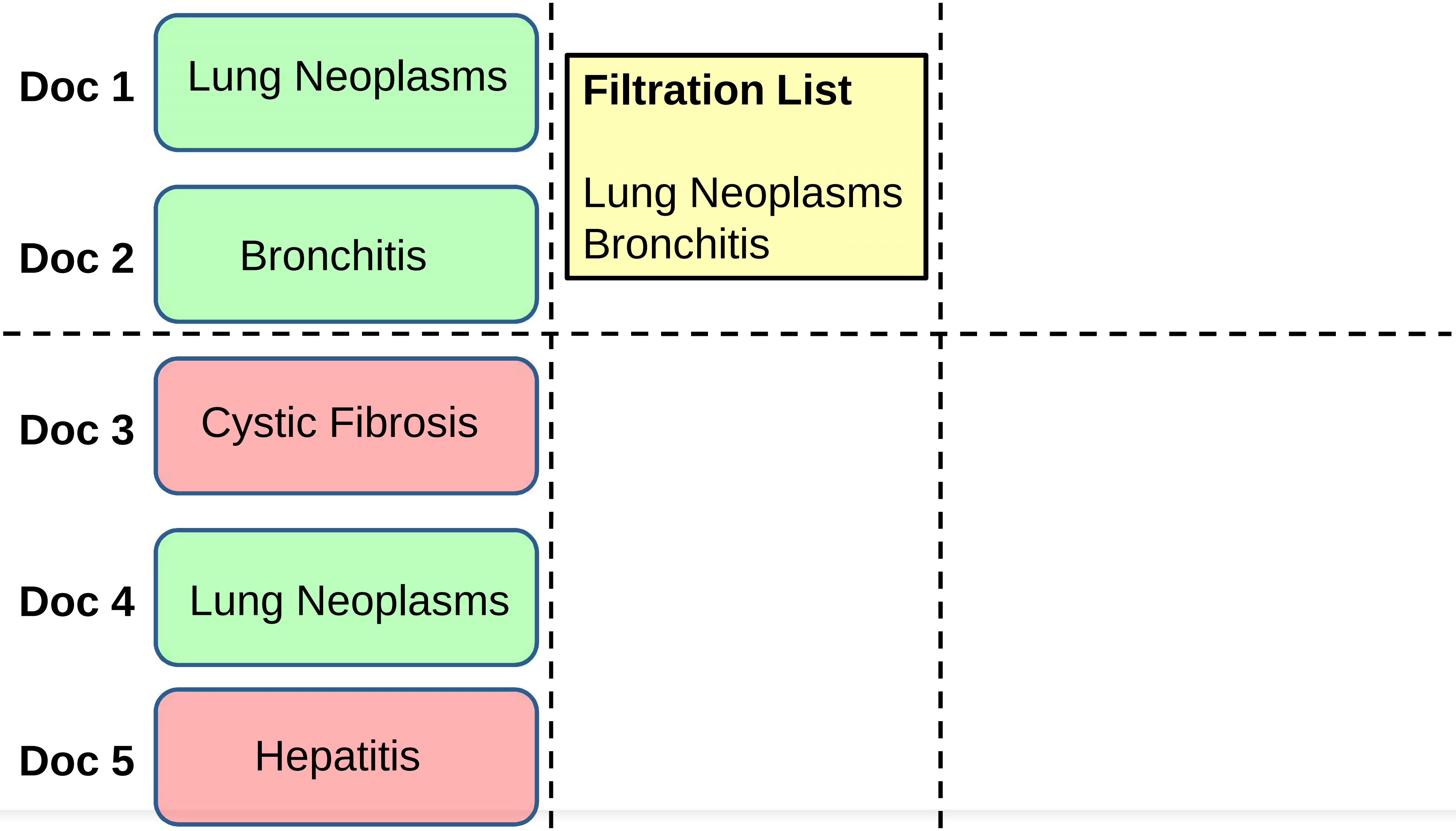
Doc 2 Bronchitis

Doc 3 Cystic Fibrosis

Doc 4 Lung Neoplasms

Doc 5 Hepatitis

Method 2: MeSH Feedback



Doc 1 Lung Neoplasms

Doc 2 Bronchitis

Doc 3 Cystic Fibrosis

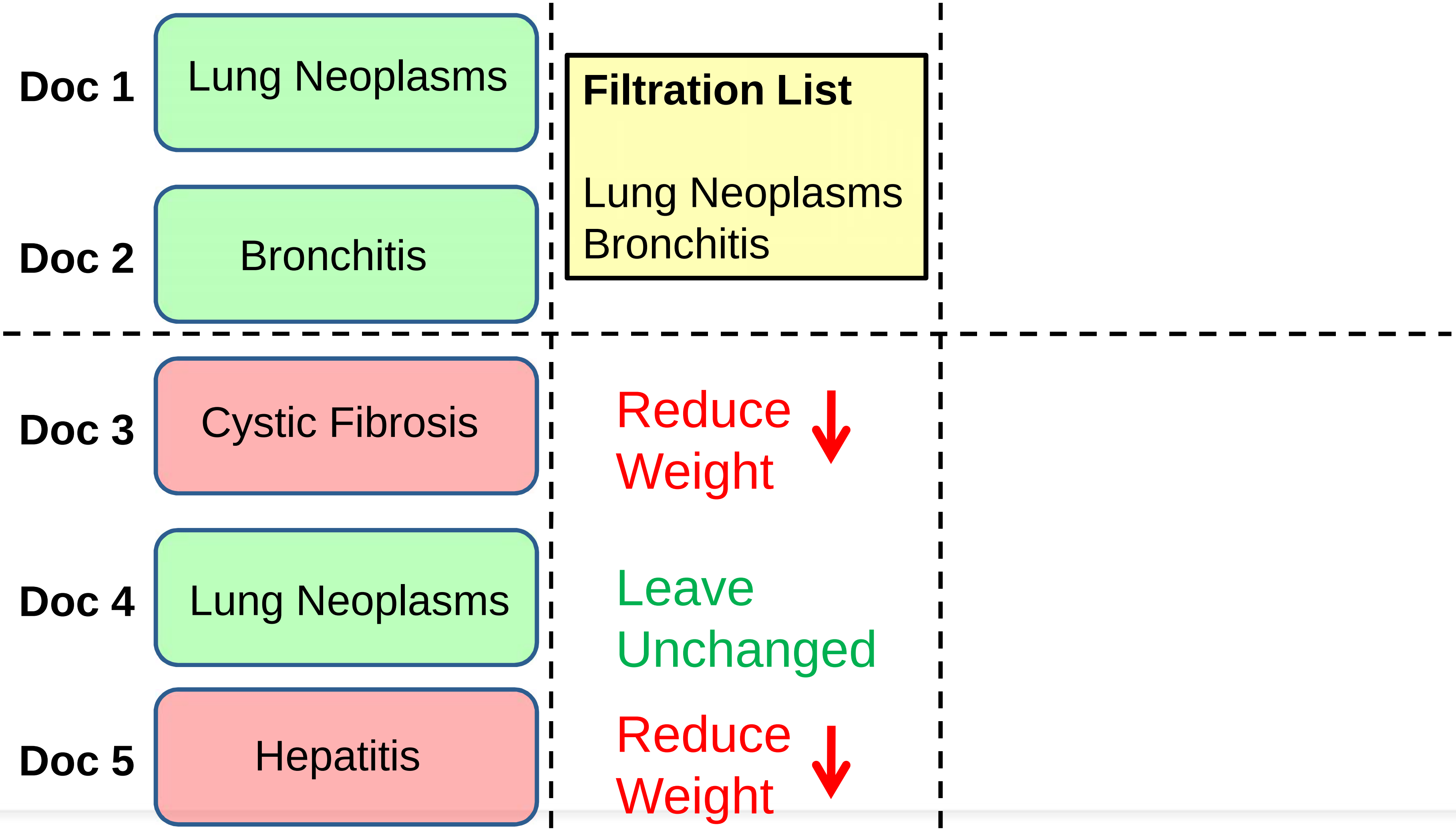
Doc 4 Lung Neoplasms

Doc 5 Hepatitis

Filtration List

Lung Neoplasms
Bronchitis

Method 2: MeSH Feedback



Doc 1 Lung Neoplasms

Doc 2 Bronchitis

Filtration List

Lung Neoplasms
Bronchitis

Doc 3 Cystic Fibrosis

Reduce Weight ↓

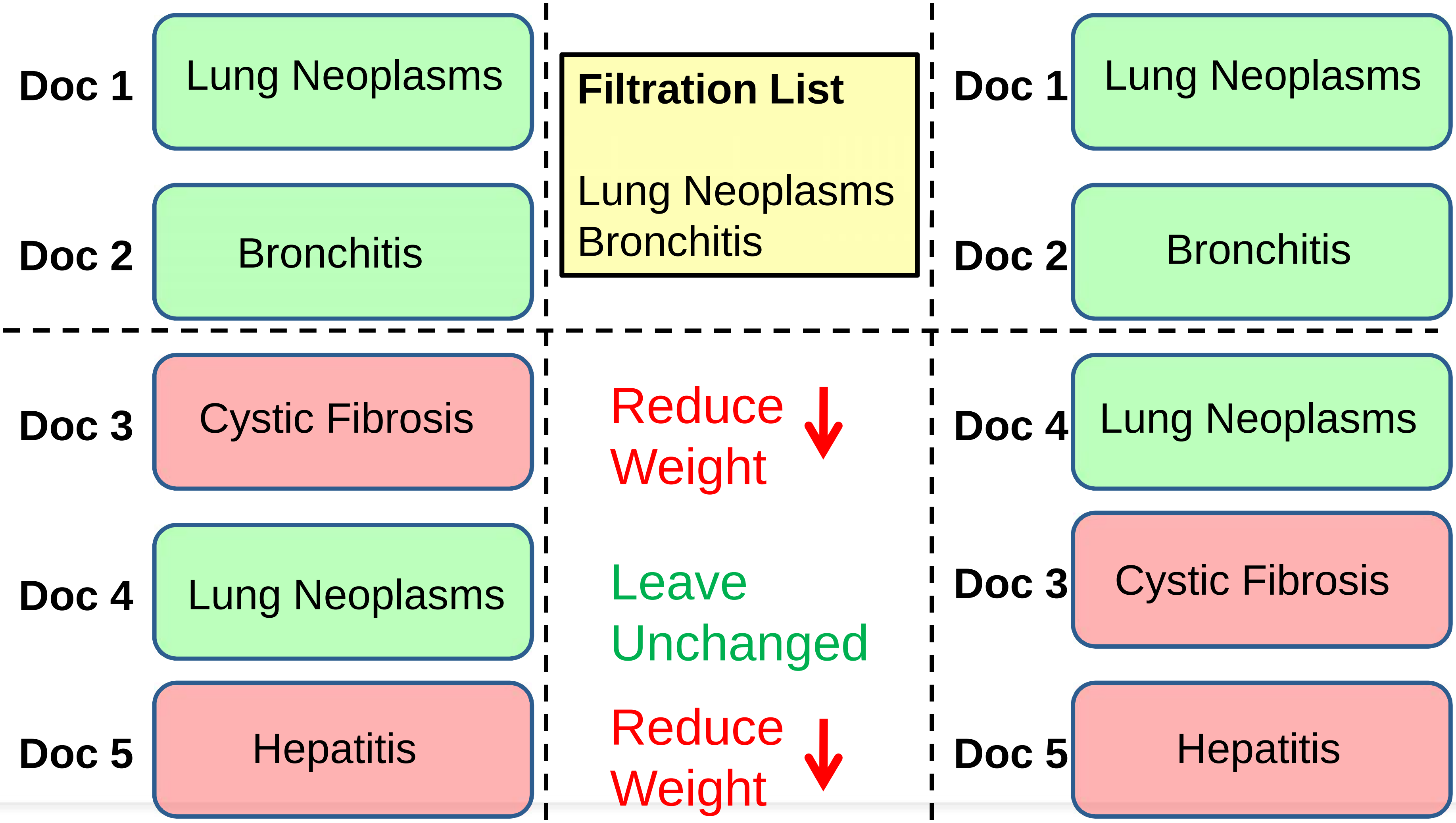
Doc 4 Lung Neoplasms

Leave Unchanged

Doc 5 Hepatitis

Reduce Weight ↓

Method: MeSH Feedback



Doc 1 Lung Neoplasms

Doc 2 Bronchitis

Doc 3 Cystic Fibrosis

Doc 4 Lung Neoplasms

Doc 5 Hepatitis

Filtration List

Lung Neoplasms
Bronchitis

Reduce Weight ↓

Leave Unchanged

Reduce Weight ↓

Doc 1 Lung Neoplasms

Doc 2 Bronchitis

Doc 4 Lung Neoplasms

Doc 3 Cystic Fibrosis

Doc 5 Hepatitis

Retrieval Results

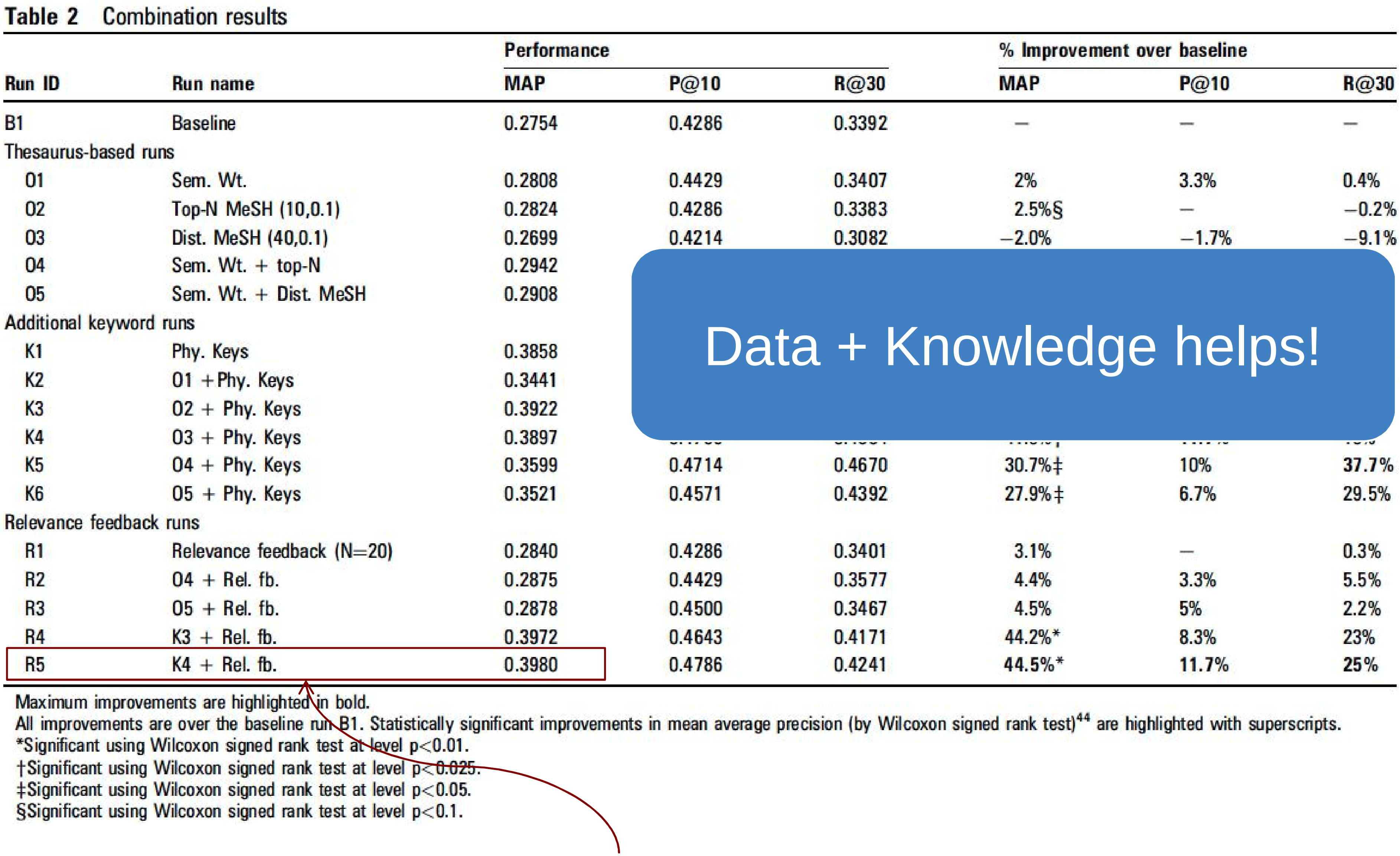


Table 2 Combination results

Run ID	Run name	Performance			% Improvement over baseline		
		MAP	P@10	R@30	MAP	P@10	R@30
B1	Baseline	0.2754	0.4286	0.3392	–	–	–
Thesaurus-based runs							
O1	Sem. Wt.	0.2808	0.4429	0.3407	2%	3.3%	0.4%
O2	Top-N MeSH (10,0.1)	0.2824	0.4286	0.3383	2.5%§	–	−0.2%
O3	Dist. MeSH (40,0.1)	0.2699	0.4214	0.3082	−2.0%	−1.7%	−9.1%
O4	Sem. Wt. + top-N	0.2942	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
O5	Sem. Wt. + Dist. MeSH	0.2908	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Additional keyword runs							
K1	Phy. Keys	0.3858	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
K2	O1 +Phy. Keys	0.3441	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
K3	O2 + Phy. Keys	0.3922	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
K4	O3 + Phy. Keys	0.3897	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
K5	O4 + Phy. Keys	0.3599	0.4714	0.4670	30.7%‡	10%	37.7%
K6	O5 + Phy. Keys	0.3521	0.4571	0.4392	27.9%‡	6.7%	29.5%
Relevance feedback runs							
R1	Relevance feedback (N=20)	0.2840	0.4286	0.3401	3.1%	–	0.3%
R2	O4 + Rel. fb.	0.2875	0.4429	0.3577	4.4%	3.3%	5.5%
R3	O5 + Rel. fb.	0.2878	0.4500	0.3467	4.5%	5%	2.2%
R4	K3 + Rel. fb.	0.3972	0.4643	0.4171	44.2%*	8.3%	23%
R5	K4 + Rel. fb.	0.3980	0.4786	0.4241	44.5%*	11.7%	25%

Maximum improvements are highlighted in bold.
All improvements are over the baseline run B1. Statistically significant improvements in mean average precision (by Wilcoxon signed rank test)[44] are highlighted with superscripts.
*Significant using Wilcoxon signed rank test at level p<0.01.
†Significant using Wilcoxon signed rank test at level p<0.025.
‡Significant using Wilcoxon signed rank test at level p<0.05.
§Significant using Wilcoxon signed rank test at level p<0.1.

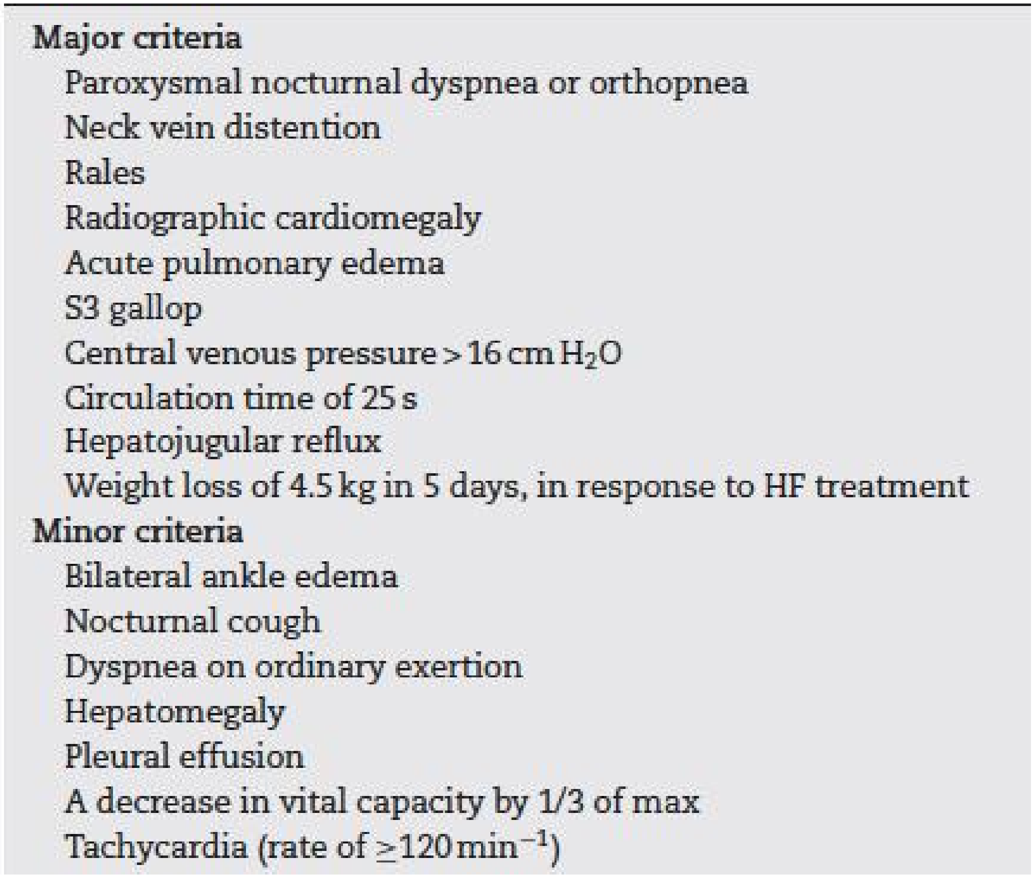
Data + Knowledge helps!

Best performing run

CASE 2: HEART FAILURE SIGNS AND SYMPTOMS

Roy J. Byrd, Steven R. Steinhubl, Jimeng Sun, Shahram Ebadollahi, Walter F. Stewart. Automatic identification of heart failure diagnostic criteria, using text analysis of clinical notes from electronic health records. International Journal of Medical Informatics 2013

Framingham HF Signs and Symptoms



Major criteria

- Paroxysmal nocturnal dyspnea or orthopnea
- Neck vein distention
- Rales
- Radiographic cardiomegaly
- Acute pulmonary edema
- S3 gallop
- Central venous pressure > 16 cm H_2O
- Circulation time of 25 s
- Hepatojugular reflux
- Weight loss of 4.5 kg in 5 days, in response to HF treatment

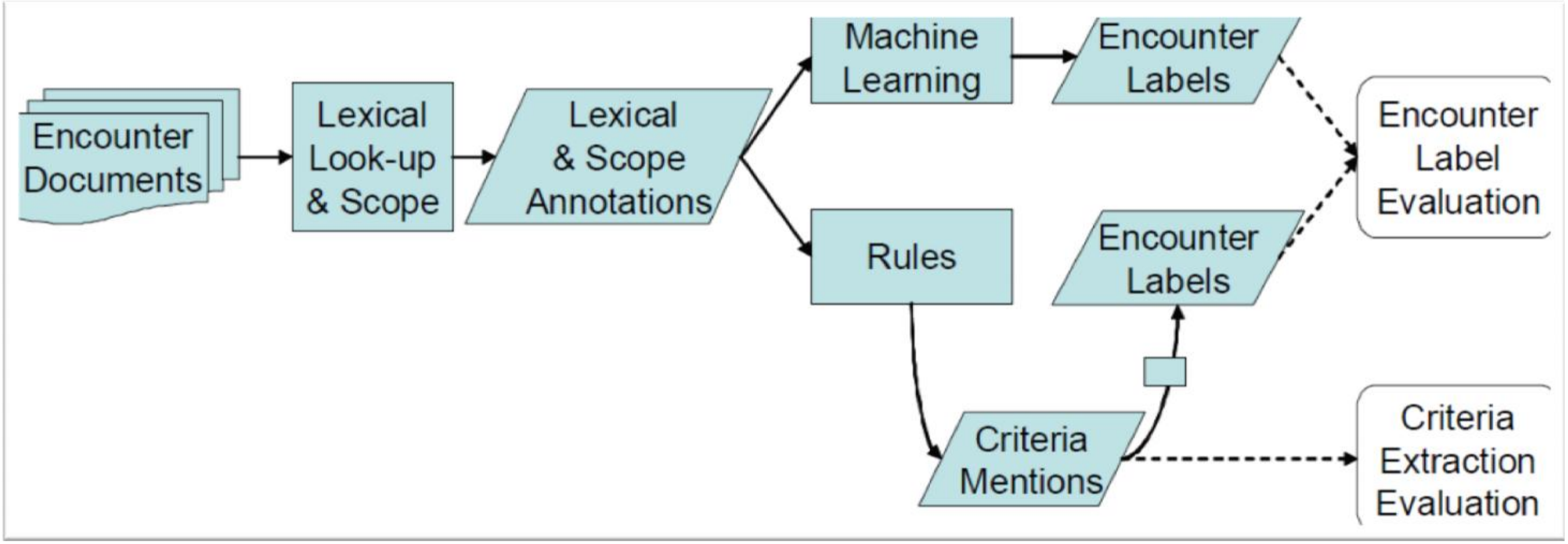
Minor criteria

- Bilateral ankle edema
- Nocturnal cough
- Dyspnea on ordinary exertion
- Hepatomegaly
- Pleural effusion
- A decrease in vital capacity by 1/3 of max
- Tachycardia (rate of $\geq 120 \text{ min}^{-1}$)

- Framingham criteria for HF* are signs and symptoms that are documented even at primary care visits

* McKee PA, Castelli WP, McNamara PM, Kannel WB. The natural history of congestive heart failure: the Framingham study. N Engl J Med. 1971;**285**(26):1441-6.

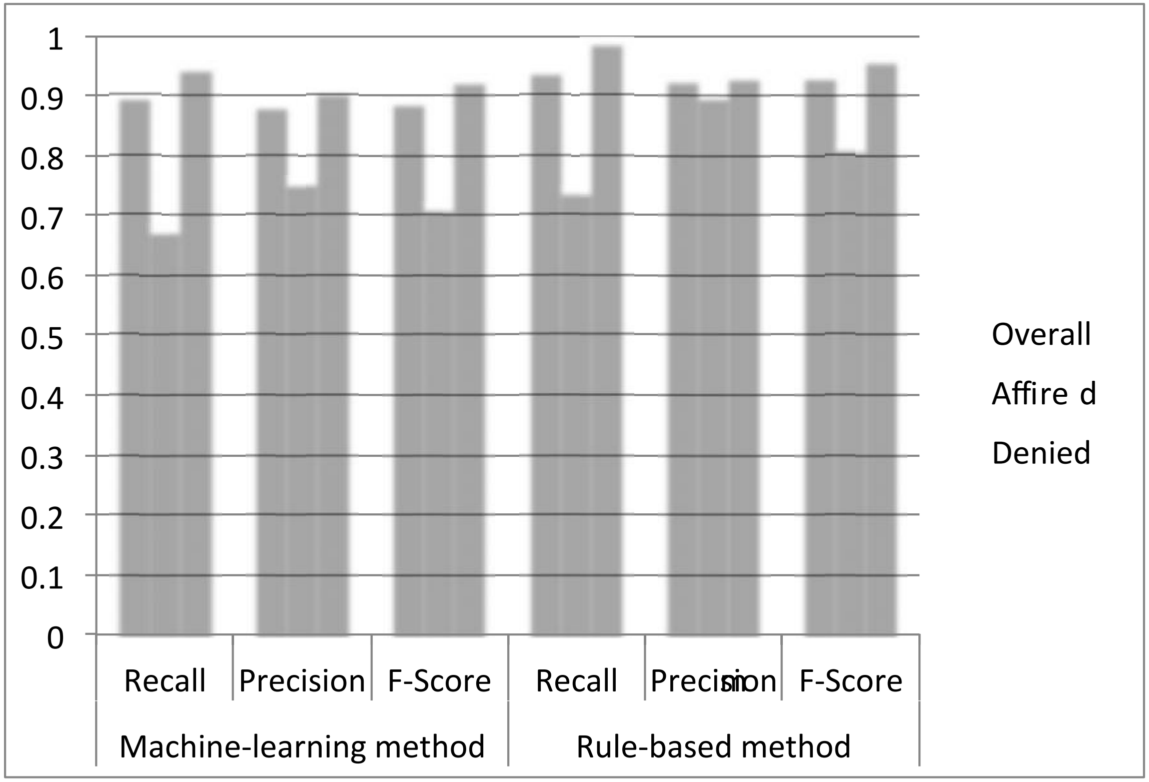
Natural Language Processing (NLP) Pipeline



- Criteria extraction comes from sentence level.
- Encounter label comes from the entire note.

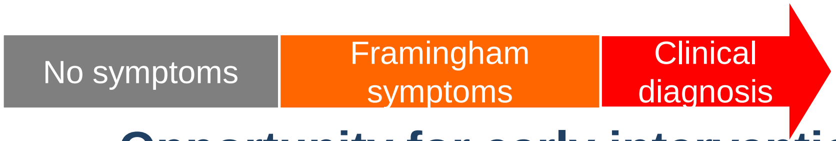
Roy J. Byrd, Steven R. Steinhubl, Jimeng Sun, Shahram Ebadollahi, Walter F. Stewart. Automatic identification of heart failure diagnostic criteria, using text analysis of clinical notes from electronic health records. International Journal of Medical Informatics 2013

Performance on Encounter Level on Test Set



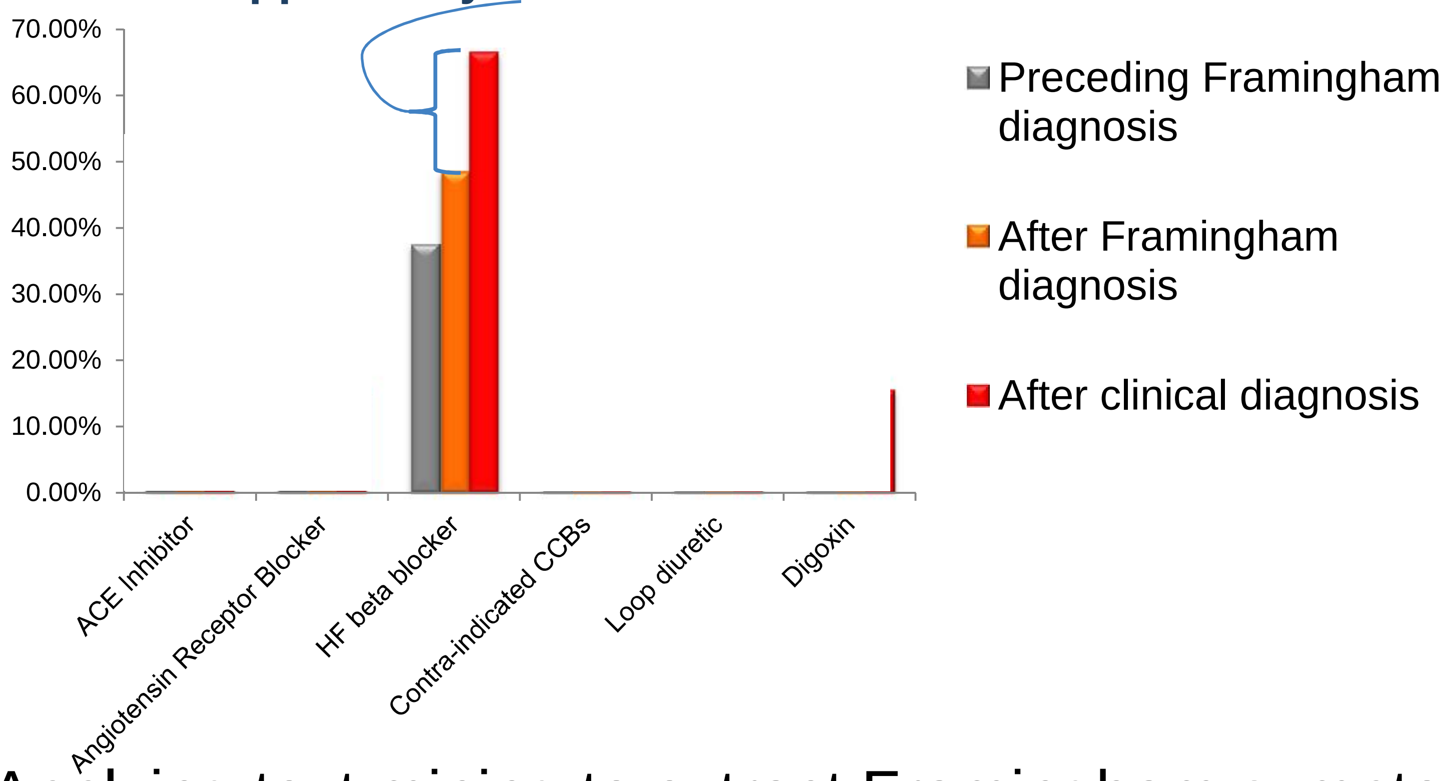
- Machine learning method: decision tree
- Rule-based method is to construct grammars by computational linguists
- Manually constructed rules are more accurate but more effort to construct than automatic rules from learning a decision tree

Potential Impact on Evidence-based Therapies


No symptoms
Framingham symptoms
Clinical diagnosis

3,168 patients eventually all diagnosed with HF

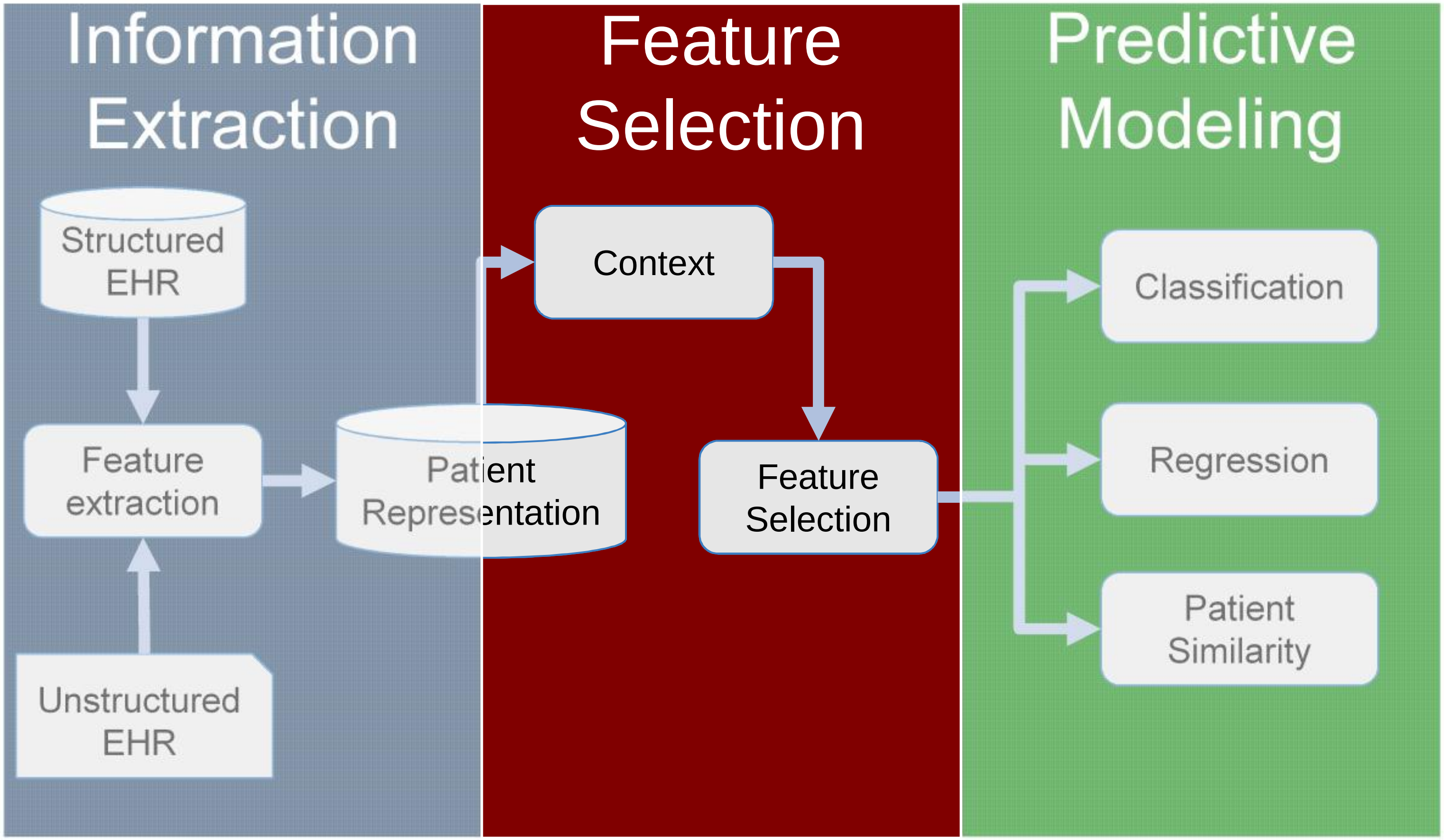
Opportunity for early intervention



- Applying text mining to extract Framingham symptoms can help trigger early intervention

Vhavakrishnan R, Steinhubl SR, Sun J, et al. Potential impact of predictive models for early detection of heart failure on the initiation of evidence-based therapies. J Am Coll Cardiol. 2012;59(13s1):E949-E949.

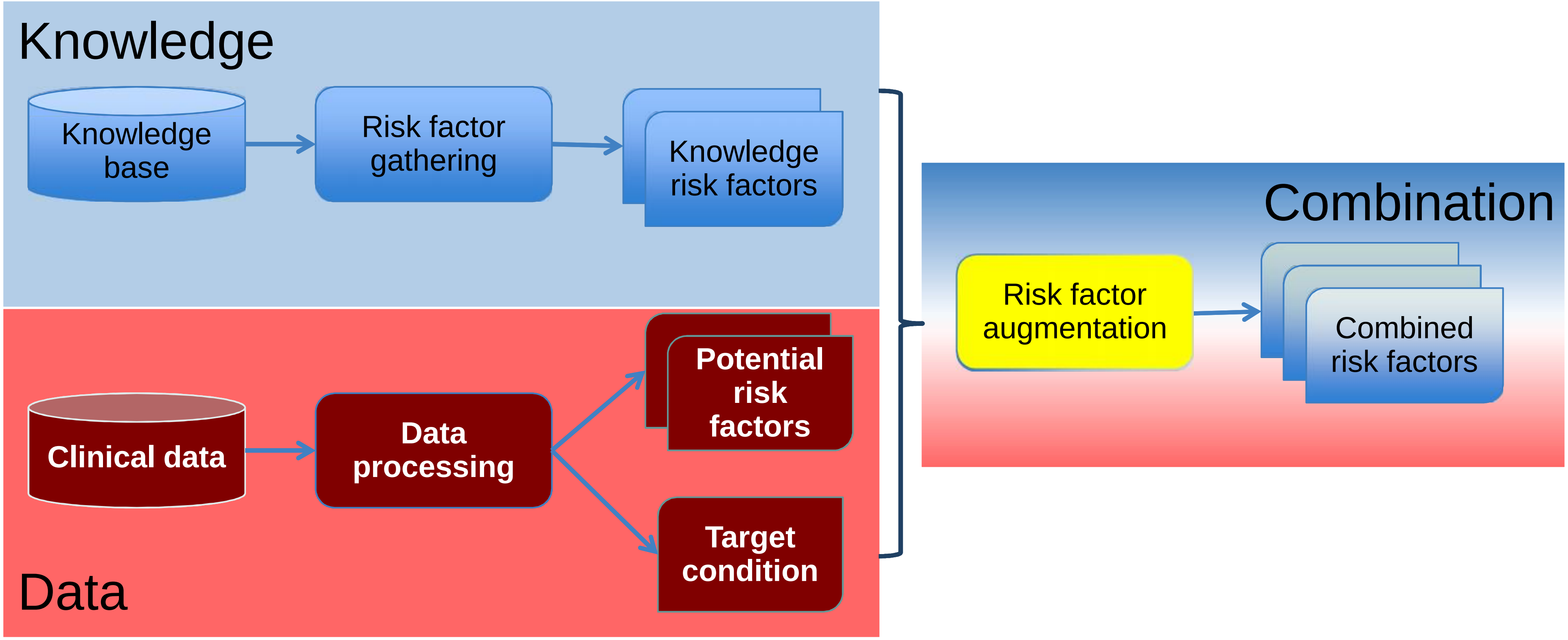
Analytic Platform



KNOWLEDGE+DATA FEATURE SELECTION

Jimeng Sun, Jianying Hu, Dijun Luo, Marianthi Markatou, Fei Wang, Shahram Ebadollahi, Steven E. Steinhubl, Zahra Daar, Walter F. Stewart. Combining Knowledge and Data Driven Insights for Identifying Risk Factors using Electronic Health Records. AMIA (2012).

Combining Knowledge- and Data-driven Risk Factors



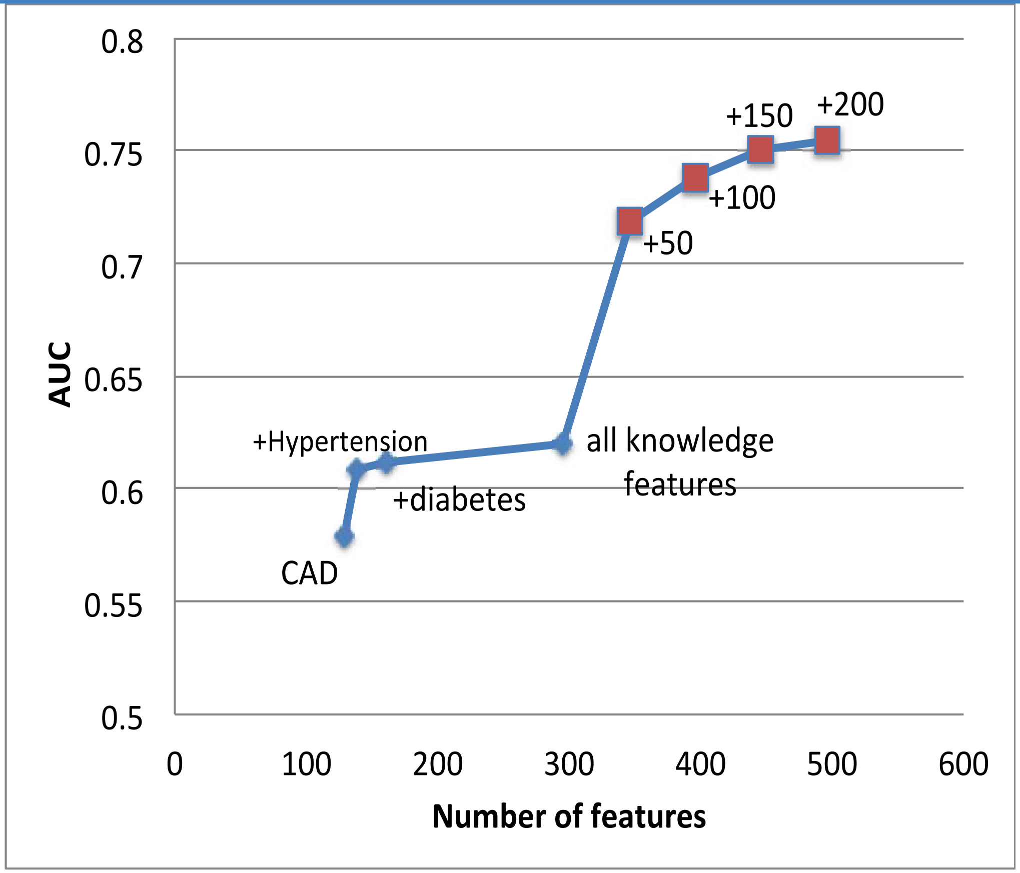
Risk Factor Augmentation

- Model Accuracy:
  - – The selected risk factors are highly predictive of the target condition
  - – Sparse feature selection through L_1 regularization
- Minimal Correlations:
  - – Between data driven risk factors and knowledge driven risk factors
  - – Among the data driven risk factors

$$f(\boldsymbol{\alpha}) = \underbrace{\frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}\|^2}_{\text{Model error}} + \frac{\beta}{4}\left[\underbrace{\sum_{i=1}^{p}\sum_{j=1}^{p}(\alpha_i\alpha_j \mathbf{x}_i{}^T\mathbf{x}_j)^2}_{\text{Correlation among data-driven features}} + \underbrace{\sum_{i=1}^{p}\sum_{j=p+1}^{p+q}(\alpha_i\alpha_j \mathbf{x}_i{}^T\mathbf{x}_j)^2}_{\text{Correlation between data- and knowledge-driven features}}\right] + \underbrace{\lambda\|\boldsymbol{\alpha}\|_1}_{\text{Sparse Penalty}}$$

Dijun Luo, Fei Wang, Jimeng Sun, Marianthi Markatou, Jianying Hu,Shahram Ebadollahi, SOR: Scalable Orthogonal Regression for Low-Redundancy Feature Selection and its Healthcare Applications. SDM’12

Prediction Results using Selected Features



- AUC significantly improves as complementary data driven risk factors are added into existing knowledge based risk factors.
- A significant AUC increase occurs when we add first 50 data driven features

Jimeng Sun, Jianying Hu, Dijun Luo, Marianthi Markatou, Fei Wang, Shahram Ebadollahi, Steven E. Steinhubl, Zahra Daar, Walter F. Stewart. Combining Knowledge and Data Driven Insights for Identifying Risk Factors using Electronic Health Records. AMIA (2012)

Clinical Validation of Data-driven Features

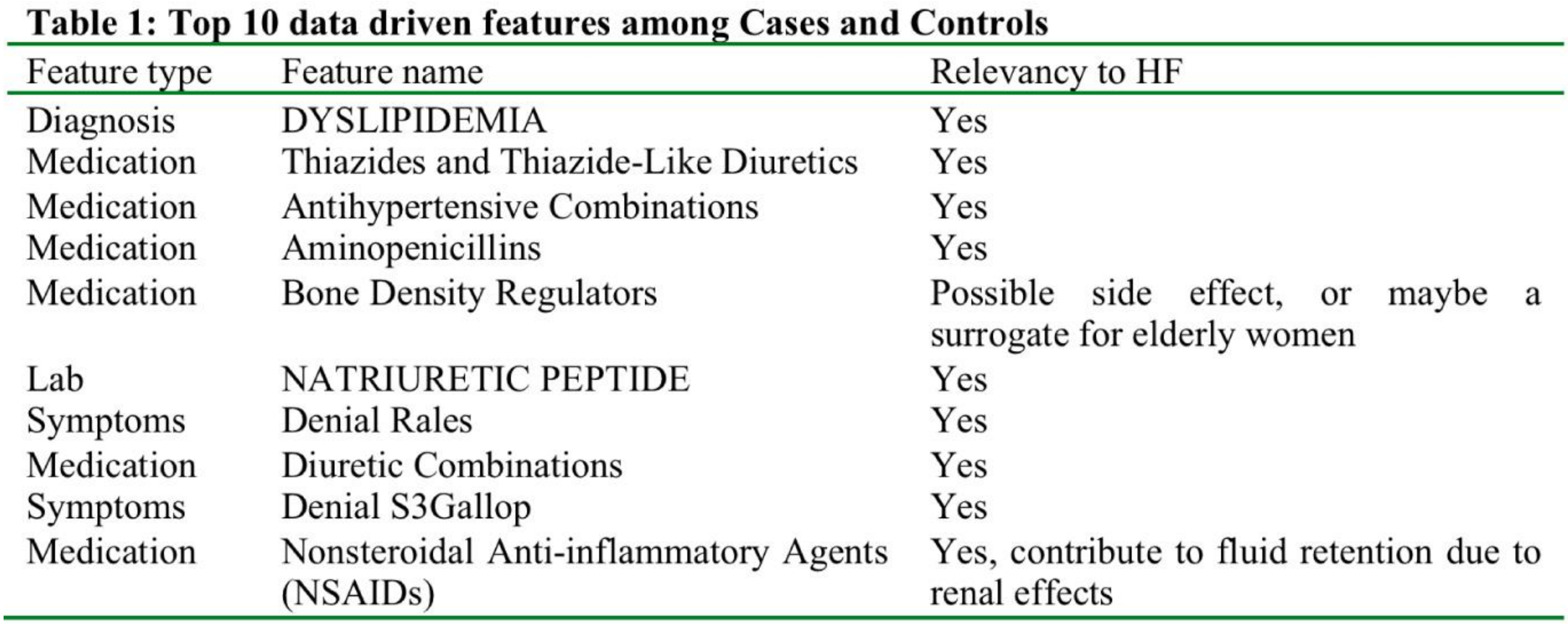
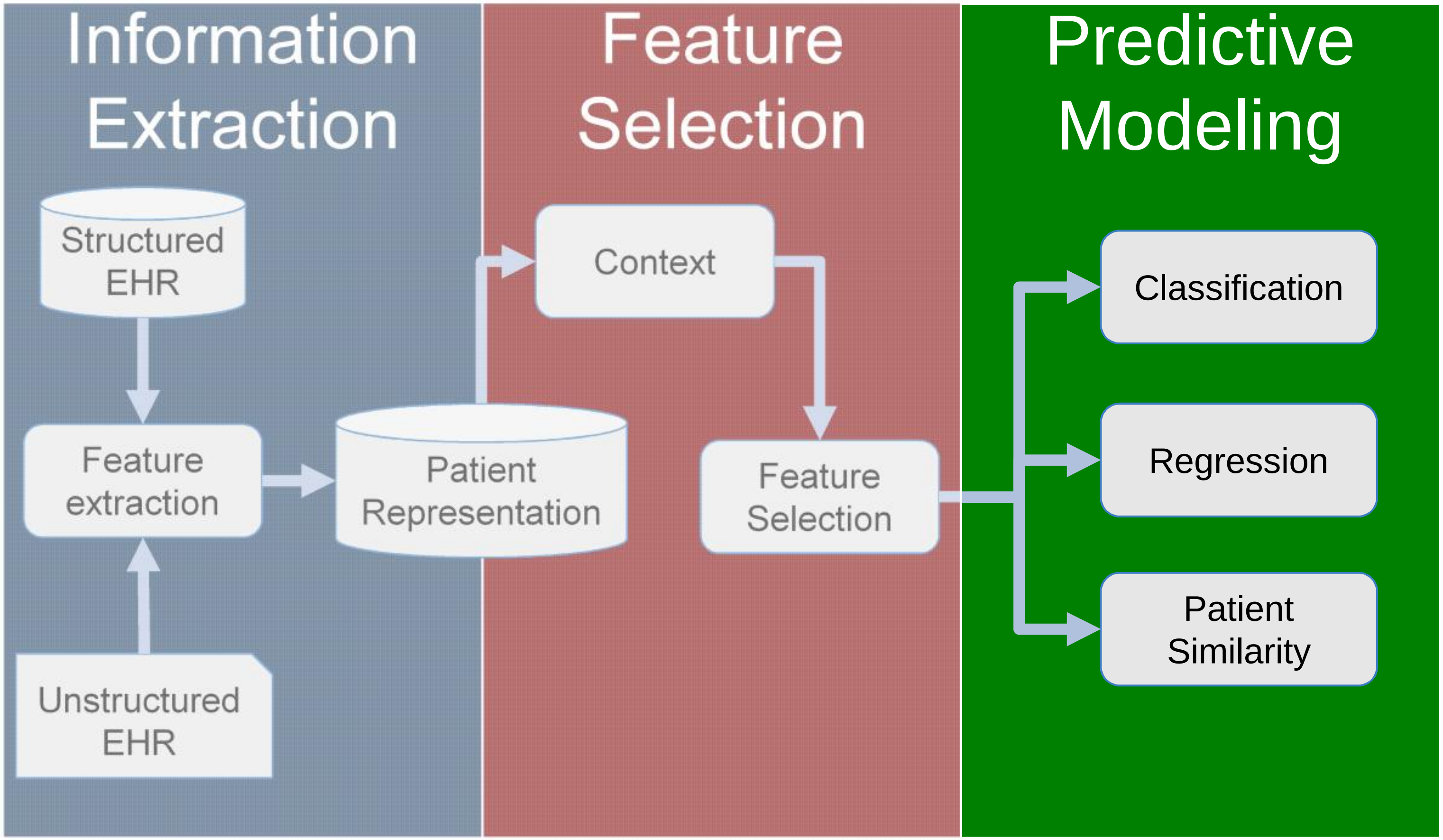


Table 1: Top 10 data driven features among Cases and Controls

Feature type	Feature name	Relevancy to HF
Diagnosis	DYSLIPIDEMIA	Yes
Medication	Thiazides and Thiazide-Like Diuretics	Yes
Medication	Antihypertensive Combinations	Yes
Medication	Aminopenicillins	Yes
Medication	Bone Density Regulators	Possible side effect, or maybe a surrogate for elderly women
Lab	NATRIURETIC PEPTIDE	Yes
Symptoms	Denial Rales	Yes
Medication	Diuretic Combinations	Yes
Symptoms	Denial S3Gallop	Yes
Medication	Nonsteroidal Anti-inflammatory Agents (NSAIDs)	Yes, contribute to fluid retention due to renal effects

- 9 out of 10 are considered relevant to HF
- The data driven features are complementary to the existing knowledge-driven features

Analytic Platform



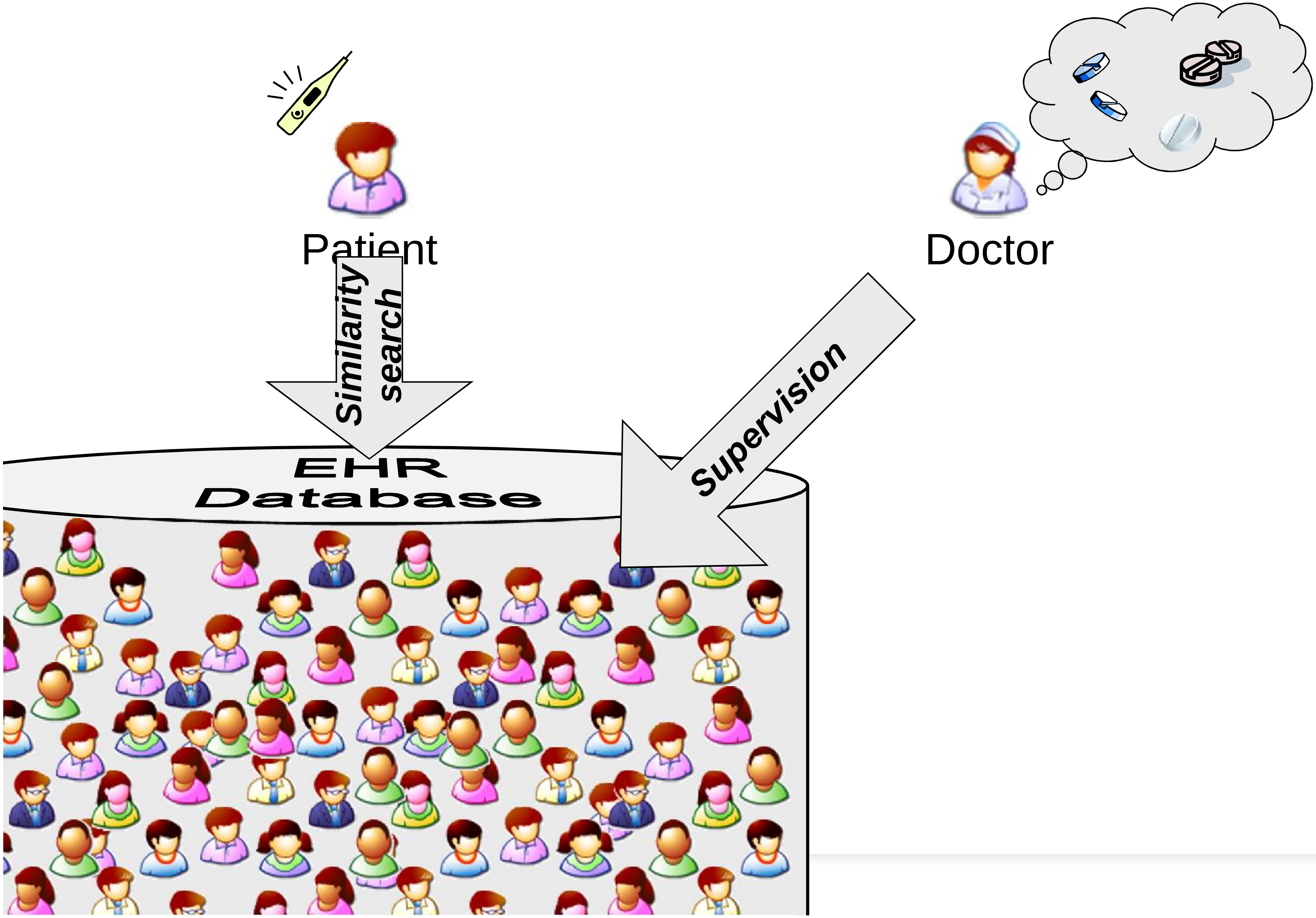
PREDICTIVE MODEL

Anatomy of Clinical Predictive Model

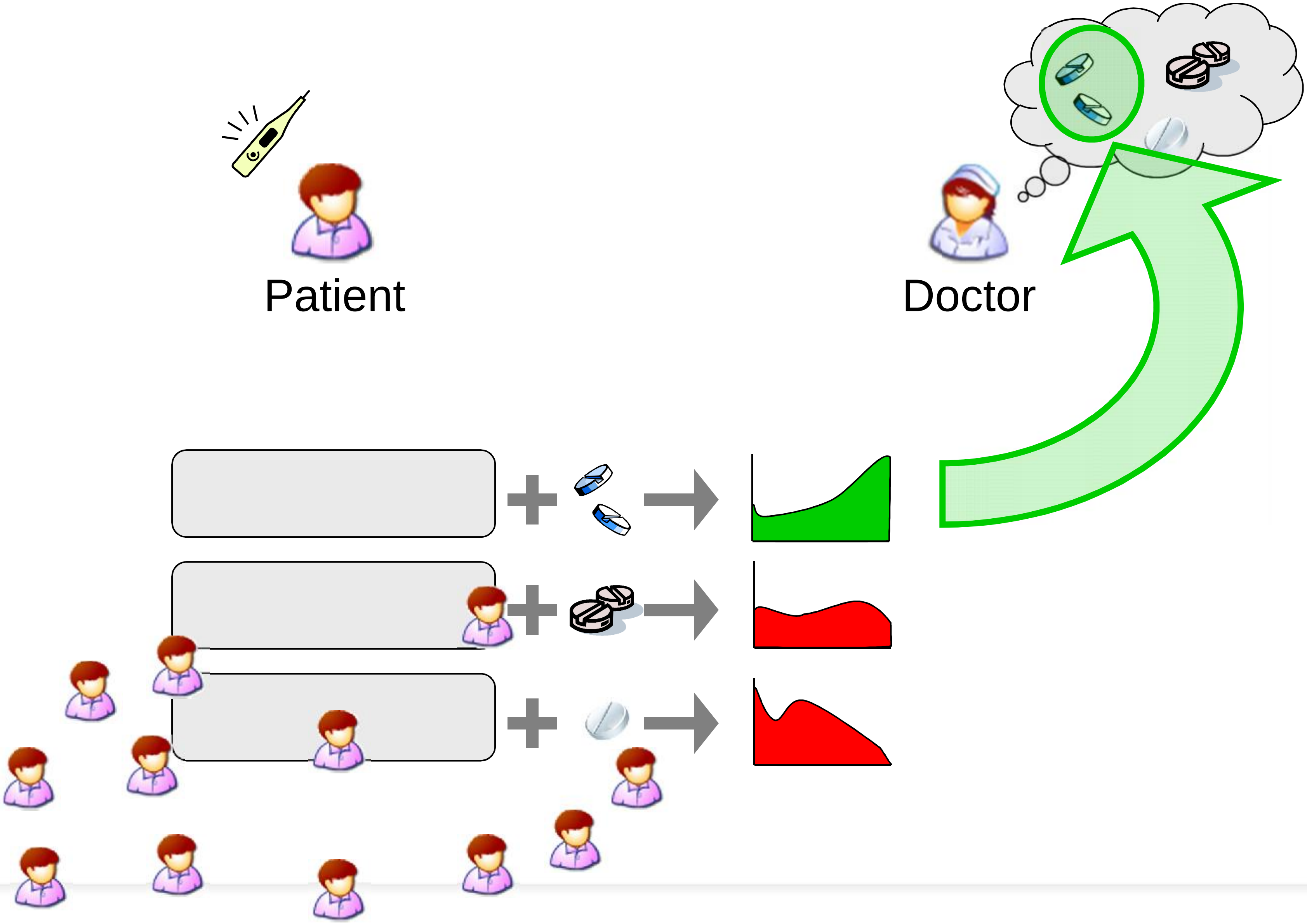
- Prediction Models
  - Continuous outcome: Regression
  - Categorical outcome: Classification
    - Logistic regression
  - Survival outcome
    - Cox Proportional Hazard Regression
  - **Patient Similarity**
- Case study: Heart failure onset prediction

PATIENT SIMILARITY

Intuition of Patient Similarity

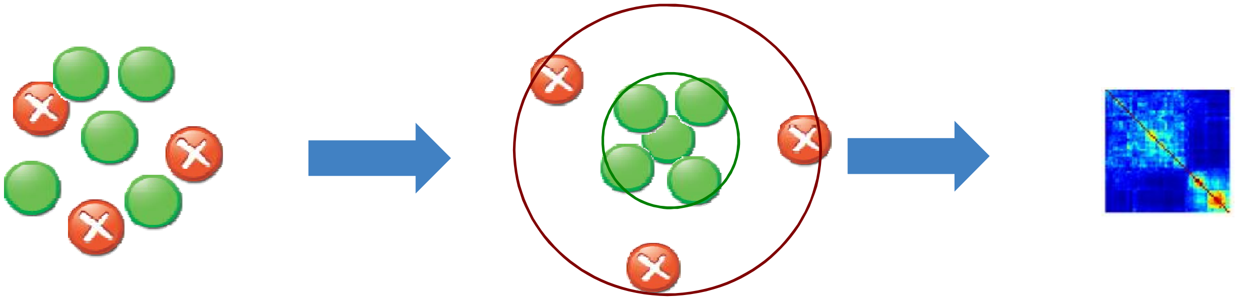


Intuition of Patient Similarity



Summary on Patient Similarity

- Patient similarity learns a customized distance metric for a specific clinical context



- Extension 1: Composite distance integration (Comdi) [SDM'11a]
  - How to jointly learn a distance by multiple parties without data sharing?
- Extension 2: Interactive metric update (iMet) [SDM'11b]
  - How to interactively update an existing distance measure?

1. Jimeng Sun, Fei Wang, Jianying Hu, Shahram Edabollahi: Supervised patient similarity measure of heterogeneous patient records. SIGKDD Explorations 14(1): 16-24 (2012)
2. Fei Wang, Jimeng Sun, Shahram Ebadollahi: Integrating Distance Metrics Learned from Multiple Experts and its Application in Inter-Patient Similarity Assessment. SDM 2011: 59-70 56
3. Fei Wang, Jimeng Sun, Jianying Hu, Shahram Ebadollahi: iMet: Interactive Metric Learning in Healthcare Applications. SDM 2011: 944-955

CASE STUDY: HEART FAILURE PREDICTION

Motivations for Early Detection of Heart Failure

- Heart failure (HF) is a complex disease
- Huge Societal Burden

5 millions HF patients in US

0.5 millions new cases each year

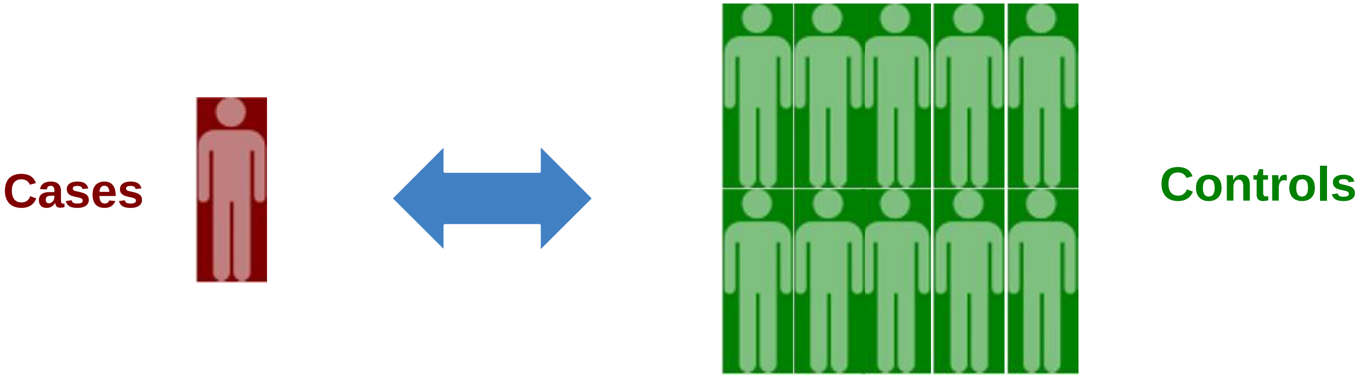
20% life time risk after 40 year old

48% 5 year mortality rate

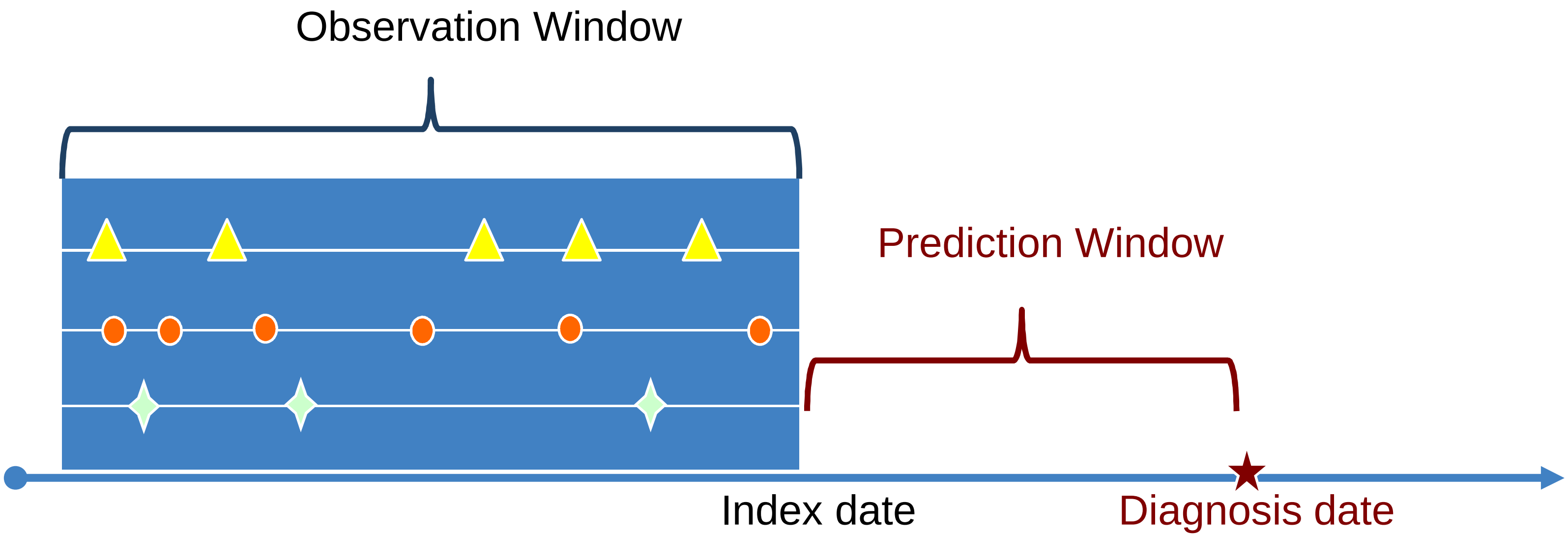
- For payers
  - Reduce cost and hospitalization
  - Improve the existing clinical guidance of HF prevention
- For providers
  - Slow or potentially reverse disease progress
  - Improve quality of life, reduce mortality

Predictive Modeling Study Design

- Goal: Classify HF cases against control patients
- Population
  - – 50,625 Patients (Geisinger Clinic PCPs)
  - – **Cases**: 4,644 case patients
  - – **Controls** 45,981 matched on age, gender and clinic



Predictive Modeling Setup



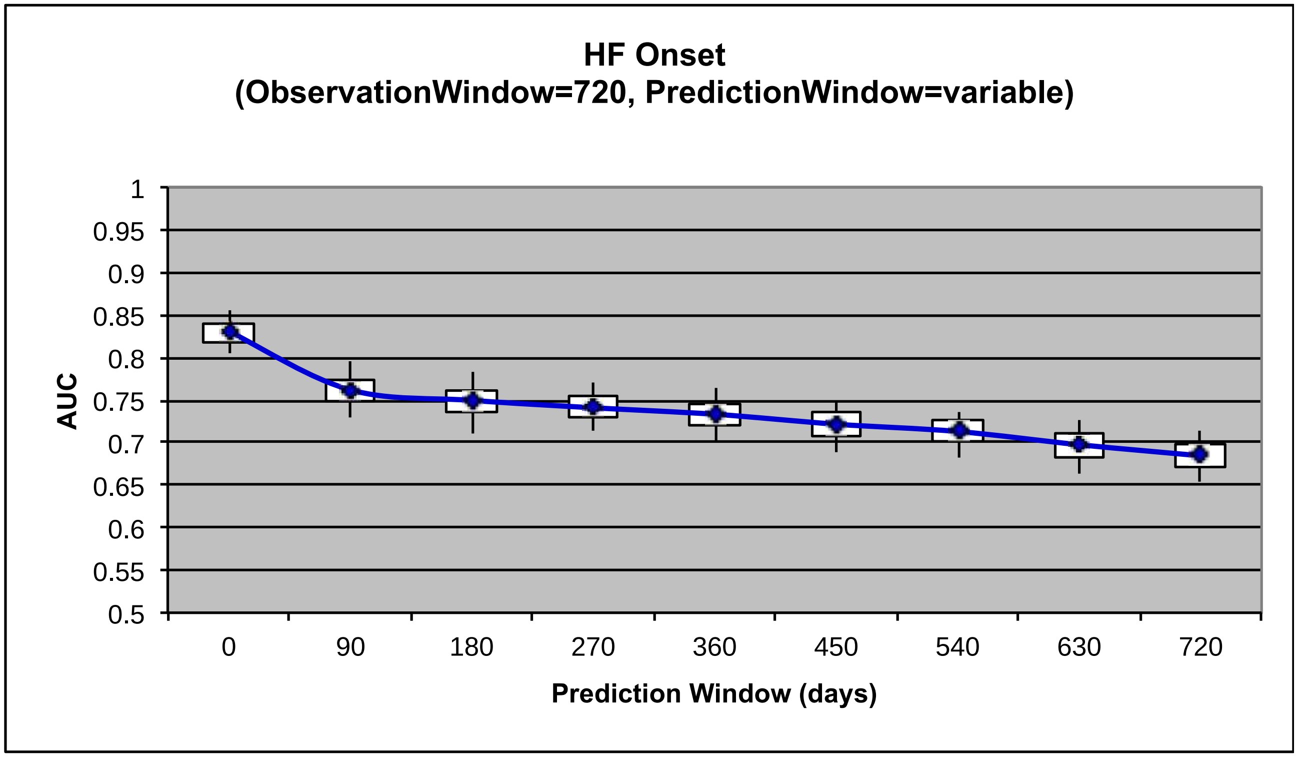
- We define
  - – Diagnosis date and index date
  - – Prediction and observation windows
- Features are constructed from the observation window and predict HF onset after the prediction window

Features

- We construct over 20K features of different types
- Through feature selection and generalization, we result in the following predictive features

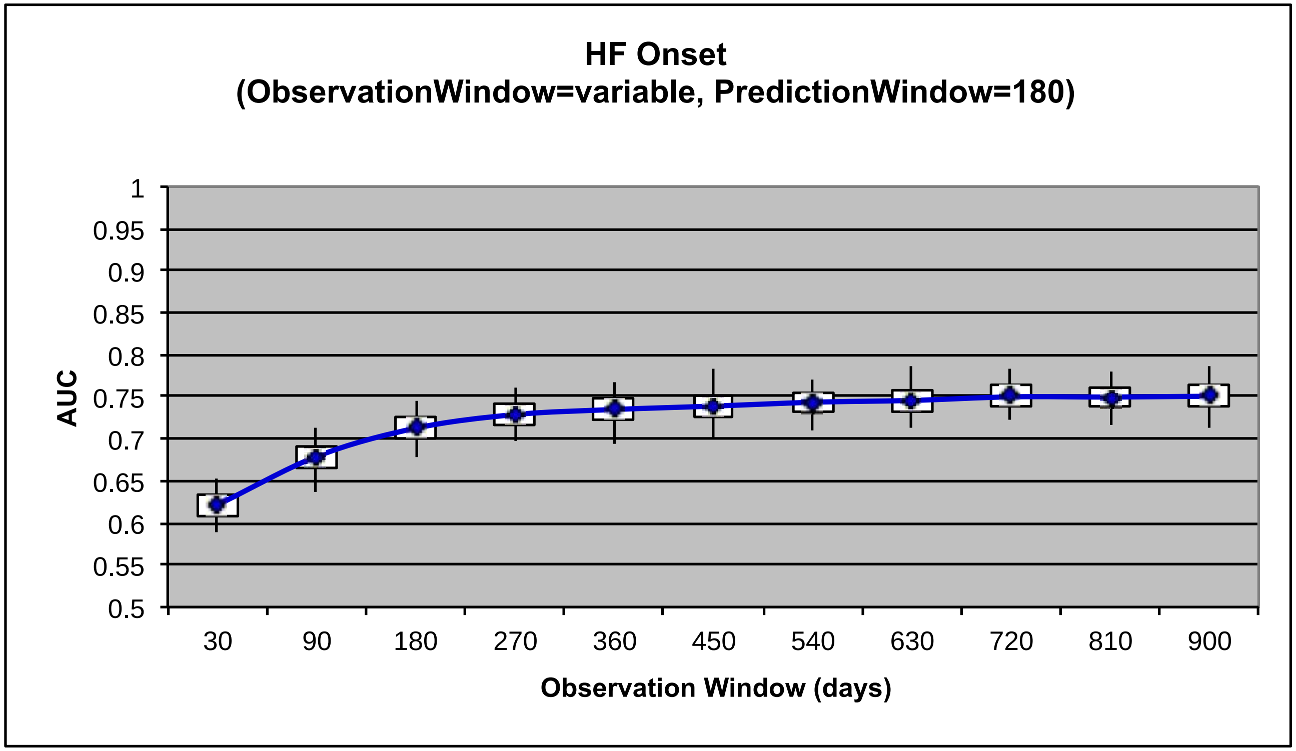
Feature type	Cardinality	Predictive Features
DIAGNOSIS	17,322	Diabetes, CHD, hypertensions, valvular disease, left ventricular hypertrophy, angina, atrial fibrillation, MI, COPD
Demographics	11	Age, race, gender, smoking status
Framingham	15	rales, cardiomegaly, acute pulmonary edema, HJReflex, ankle edema, nocturnal cough, DOExertion, hepatomegaly, pleural effusion
Lab	1,264	eGFR, LVEF, albumin, glucose, cholesterol, creatinine, cardiomegaly, heart rate, hemoglobin
Medication	3,922	antihypertensive, lipid-lowering, CCB, ACEI, ARB, beta blocker, diuretic, digitalis, antiarrhythmic
Vital	6	blood pressure and heart rate

Prediction Performance on Different Prediction Windows



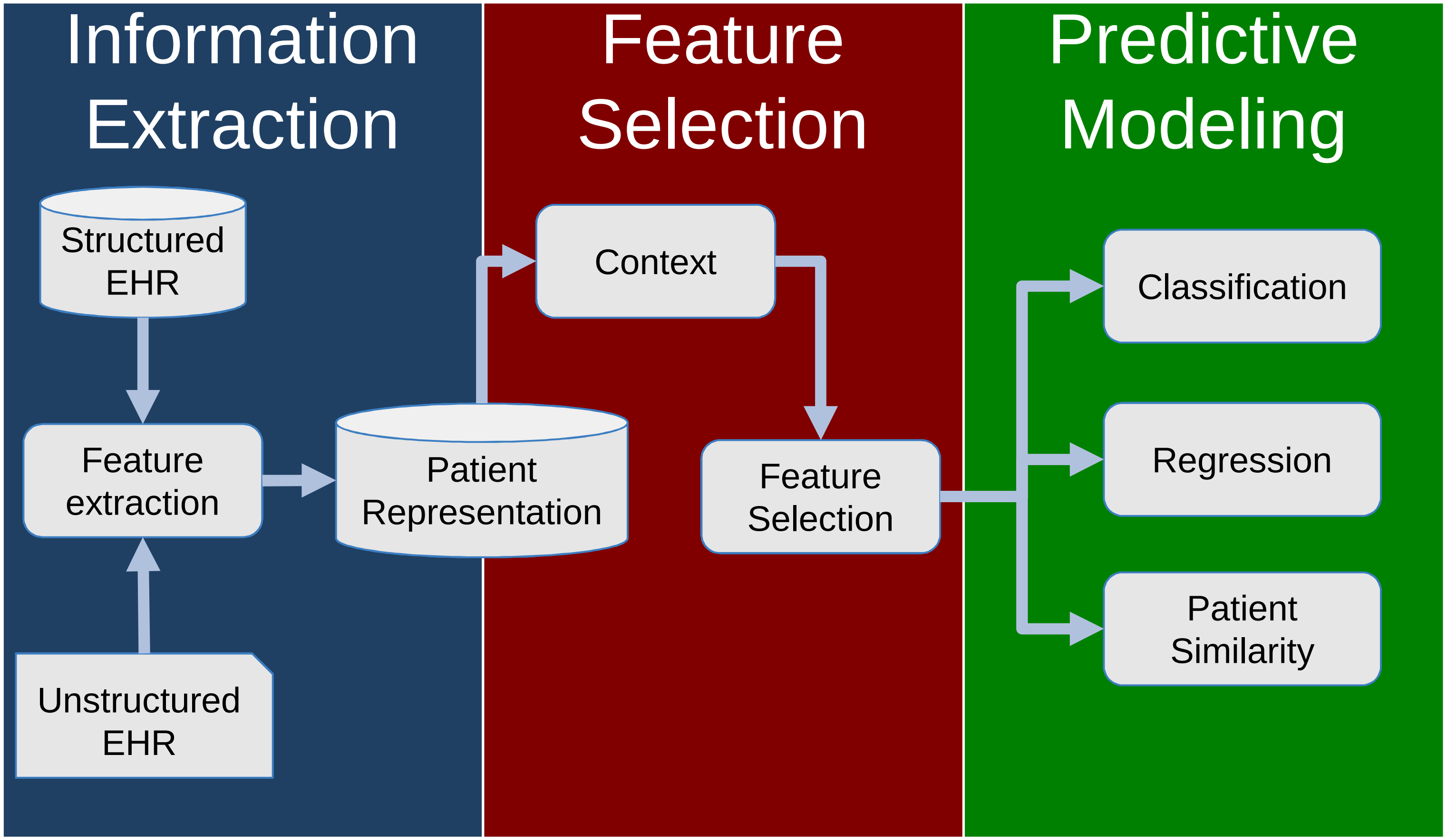
- Setting: observation window = 720 days, classifiers = random forest, evaluation mechanism = 10-fold cross-validation for 10 times
- Observation:
  - AUC slowly decreases as the prediction window increases

Prediction Performance on Different Observation Windows



- Setting: prediction window= 180 days, classifiers= random forest, evaluation mechanism =10-fold cross-validation
- Observation:
  - AUC increases as the observation window increases. i.e., more data for a longer period of time will lead to better performance of the predictive model
  - Combined features performed the best at observation window = 720 days

Analytic Platform



RESOURCES

Unstructured Clinical Data

Dataset	Link	Description
i2b2 Informatics for Integrating Biology & the Bedside	https://www.i2b2.org/NLP/DataSets/Main.php	Clinical notes used for clinical NLP challenges • 2006 Deidentification and Smoking Challenge • 2008 Obesity Challenge • 2009 Medication Challenge • 2010 Relations Challenge • 2011 Co-reference Challenge
Computational Medicine center	http://computationalmedicine.org/challenge/previous	Classifying Clinical Free Text Using Natural Language Processing

Structured EHR

Dataset	Link	Description
Texas Hospital Inpatient Discharge	http://www.dshs.state.tx.us/thcic/hospitals/Inpatientpudf.shtm	Patient: hospital location, admission type/source, claims, admit day, age, icd9 codes + surgical codes
Framingham Health Care Data Set	http://www.framinghamheartstudy.org/share/index.html	Genetic dataset for cardiovascular disease
Medicare Basic Stand Alone Claim Public Use Files	http://resdac.advantagelabs.com/cms-data/files/bsa-puf	Inpatient, skilled nursing facility, outpatient, home health agency, hospice, carrier, durable medical equipment, prescription drug event, and chronic conditions on an aggregate level
VHA Medical SAS Datasets	http://www.virec.research.va.gov/MedSAS/Overview.htm	Patient care encounters primarily for Veterans: inpatient/outpatient data from VHA facilities
Nationwide Inpatient Sample	http://www.hcup-us.ahrq.gov/nisoverview.jsp	Discharge data from 1051 hospitals in 45 states with diagnosis, procedures, status, demographics, cost, length of stay
CA Patient Discharge Data	http://www.oshpd.ca.gov/HID/Products/PatDischargeData/PublicDataSet/index.html	Discharge data for licensed general acute hospital in CA with demographic, diagnostic and treatment information, disposition, total charges
MIMIC II Clinical Database	http://mimic.physionet.org/database.html	ICU data including demographics, diagnosis, clinical measurements, lab results, interventions, notes

Thanks to Prof. Joydeep Ghosh from UT Austin for providing this information

Software

- MetaMap maps biomedical text to UMLS metathesaurus
  - Developed by NLM for parsing medical article not clinical notes
  - http://metamap.nlm.nih.gov/
- cTAKES: clinical Text Analysis and Knowledge Extraction System
  - Using Unstructured Information Management Architecture (UIMA) framework and OpenNLP toolkit
  - http://ctakes.apache.org/

Organization of this Tutorial

- **Introduction**
- **Motivating Examples**
- **Sources and Techniques for Big Data in Healthcare**
  - **Structured EHR Data**
  - **Unstructured Clinical Notes**
  - **Medical Imaging Data**
  - **Genetic Data**
  - **Other Data (Epidemiology & Behavioral)**
- **Final Thoughts and Conclusion**

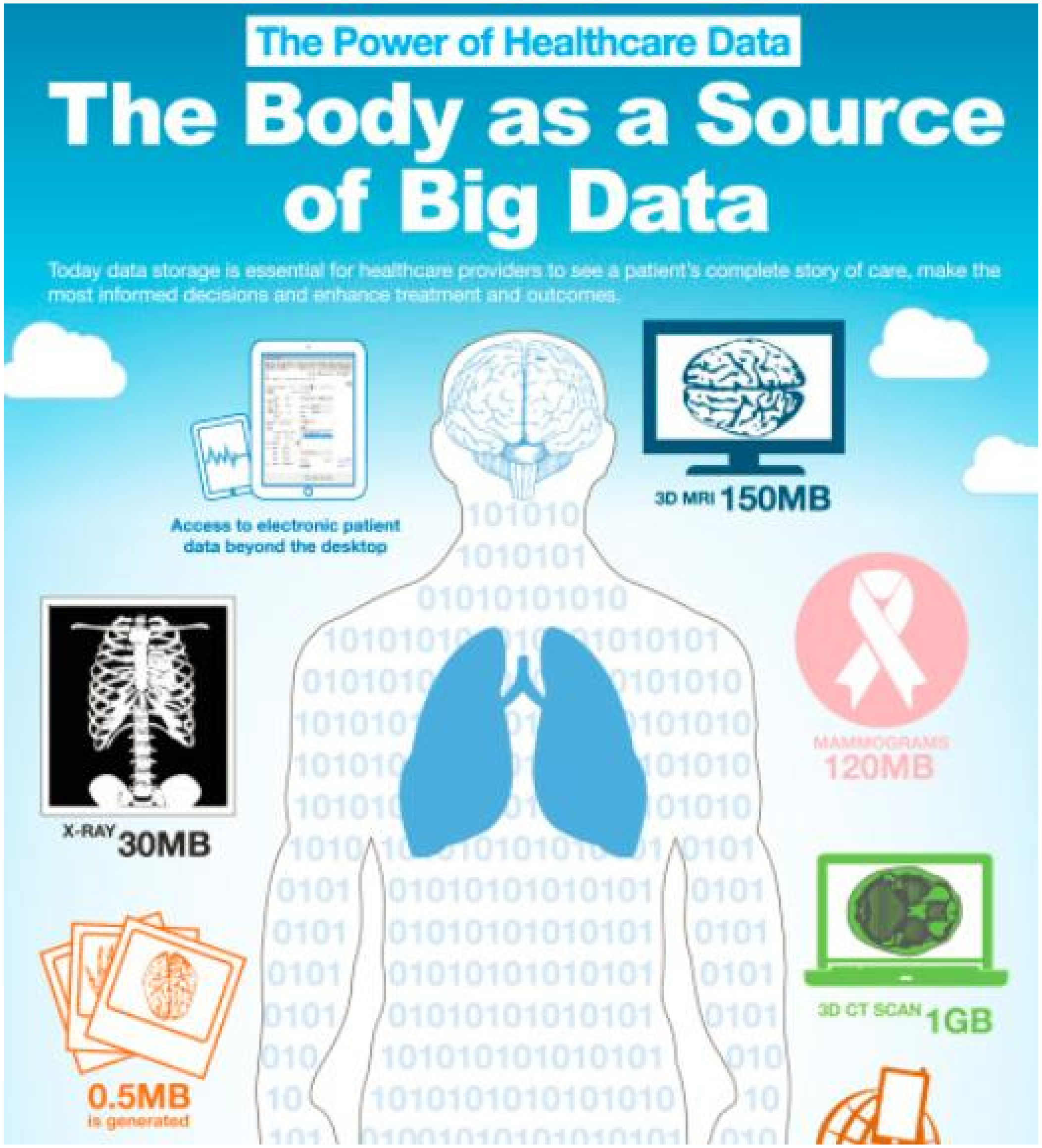
MEDICAL IMAGE DATA

Image Data is Big !!!

- By 2015, the average hospital will have two-thirds of a *petabyte* (665 terabytes) of patient data, 80% of which will be unstructured image data like CT scans and X-rays.

- Medical Imaging archives are increasing by 20%-40%

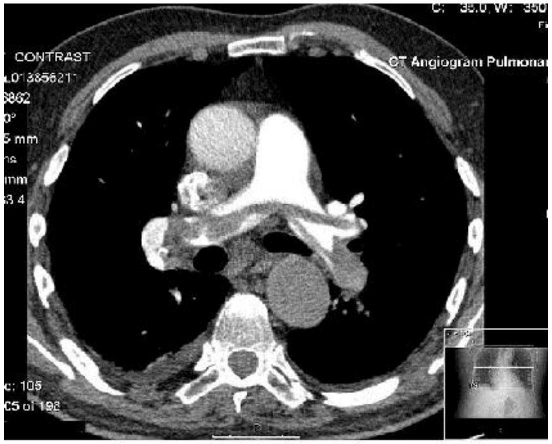
- PACS (Picture Archival & Communication Systems) system is used for storage and retrieval of the images.


The Power of Healthcare Data
The Body as a Source of Big Data
Access to electronic patient data beyond the desktop
3D MRI 150MB
X-RAY 30MB
MAMMOGRAMS 120MB
3D CT SCAN 1GB
0.5MB is generated

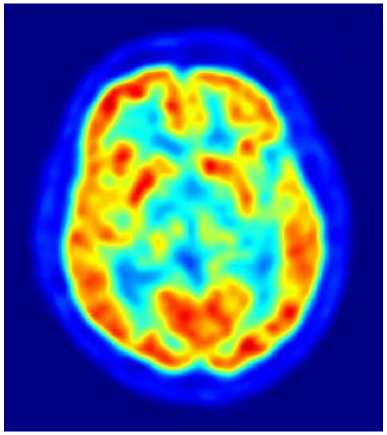
Today data storage is essential for healthcare providers to see a patient's complete story of care, make the most informed decisions and enhance treatment and outcomes.

Image Source: http://medcitynews.com/2013/03/the-body-in-bytes-medical-images-as-a-source-of-healthcare-big-data-infographic/

Popular Imaging Modalities in Healthcare Domain



Computed Tomography (CT)



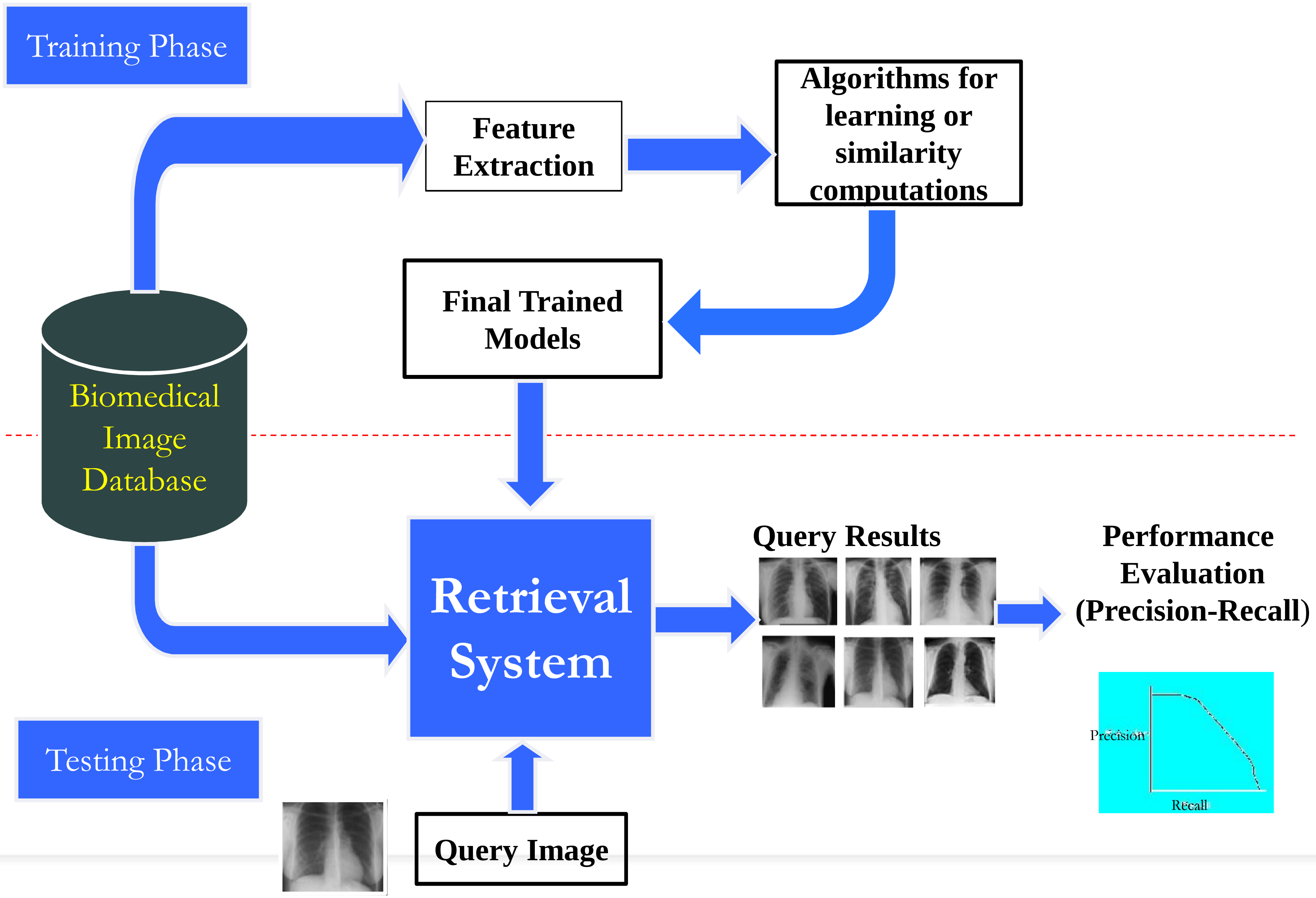
Positron Emission Tomography (PET)

Magnetic Resonance Imaging (MRI)

- The main challenge with the image data is that it is not only huge, but is also high-dimensional and complex.
- Extraction of the important and relevant features is a daunting task.
- Many research works applied image features to extract the most relevant images for a given query.

Image Source: Wikipedia

Medical Image Retrieval System



Training Phase

Feature Extraction

Algorithms for learning or similarity computations

Final Trained Models

Biomedical Image Database

Retrieval System

Query Results

Performance Evaluation (Precision-Recall)

Testing Phase

Query Image

Content-based Image Retrieval

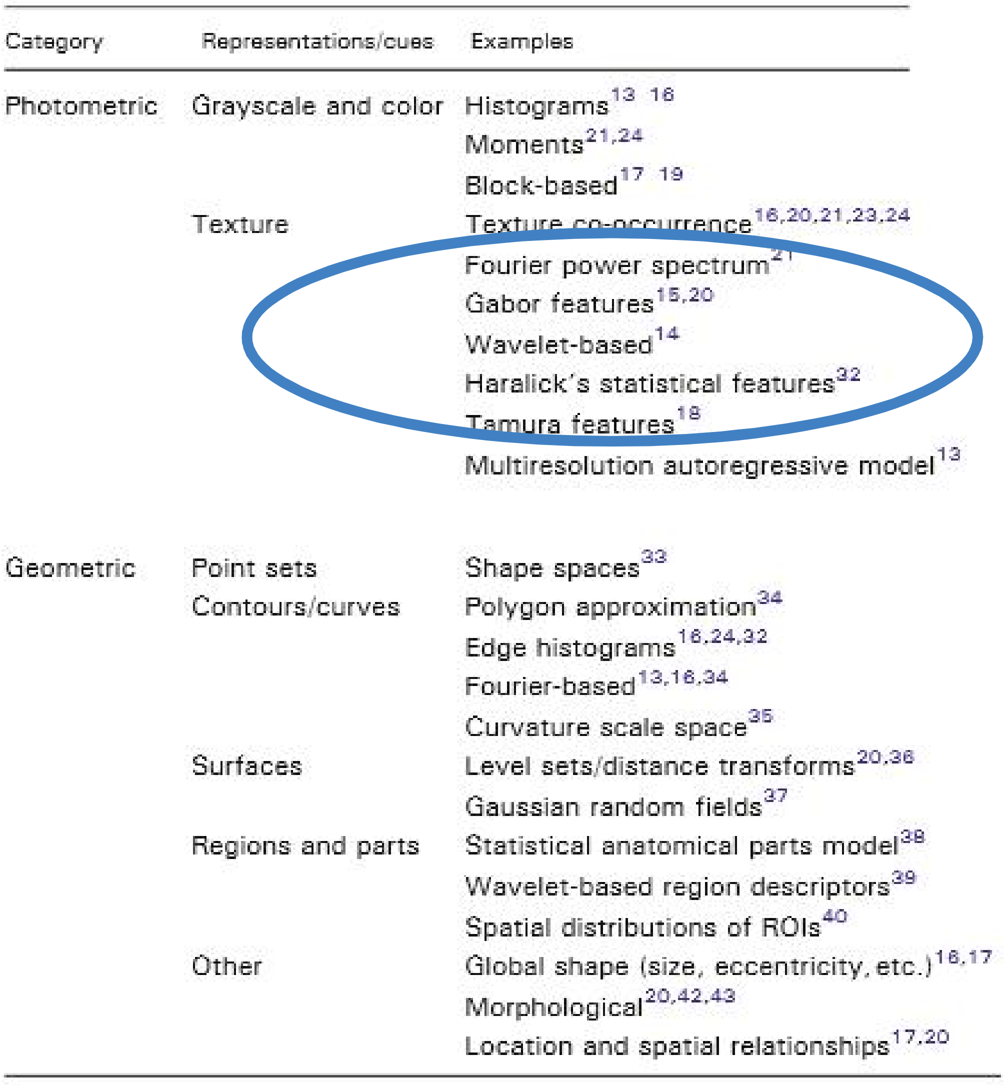
- Two components
  - Image features/descriptors - bridging the gap between the visual content and its numerical representation.
  - These representations are designed to encode color and texture properties of the image, the spatial layout of objects, and various geometric shape characteristics of perceptually coherent structures.
  - Assessment of similarities between image features based on mathematical analyses, which compare descriptors across different images.
  - Vector affinity measures such as Euclidean distance, Mahalanobis distance, KL divergence, Earth Mover's distance are amongst the widely used ones.

Medical Image Features

Photo-metric features exploit color and texture cues and they are derived directly from raw pixel intensities.

Geometric features: cues such as edges, contours, joints, polylines, and polygonal regions.

- A suitable shape representation should be extracted from the pixel intensity information by region-of interest detection, segmentation, and grouping. Due to these difficulties, geometric features are not widely used.



Category	Representations/cues	Examples
Photometric	Grayscale and color	Histograms[13 16]
		Moments[21,24]
		Block-based[17 19]
	Texture	Texture co-occurrence[16,20,21,23,24]
		Fourier power spectrum[21]
		Gabor features[15,20]
		Wavelet-based[14]
		Haralick's statistical features[32]
		Tamura features[18]
		Multiresolution autoregressive model[13]
Geometric	Point sets	Shape spaces[33]
	Contours/curves	Polygon approximation[34]
		Edge histograms[16,24,32]
		Fourier-based[13,16,34]
		Curvature scale space[35]
	Surfaces	Level sets/distance transforms[20,36]
		Gaussian random fields[37]
	Regions and parts	Statistical anatomical parts model[38]
		Wavelet-based region descriptors[39]
		Spatial distributions of ROIs[40]
	Other	Global shape (size, eccentricity, etc.)[16,17]
		Morphological[20,42,43]
		Location and spatial relationships[17,20]

Akgül, Ceyhun Burak, et al. "Content-based image retrieval in radiology: current status and future directions." Journal of Digital Imaging 24.2 (2011): 208-222.
Müller, Henning, et al. "A review of content-based image retrieval systems in medical applications-clinical benefits and future directions." *International journal of medical informatics* 73.1 (2004): 1-24.

Image CLEF Data

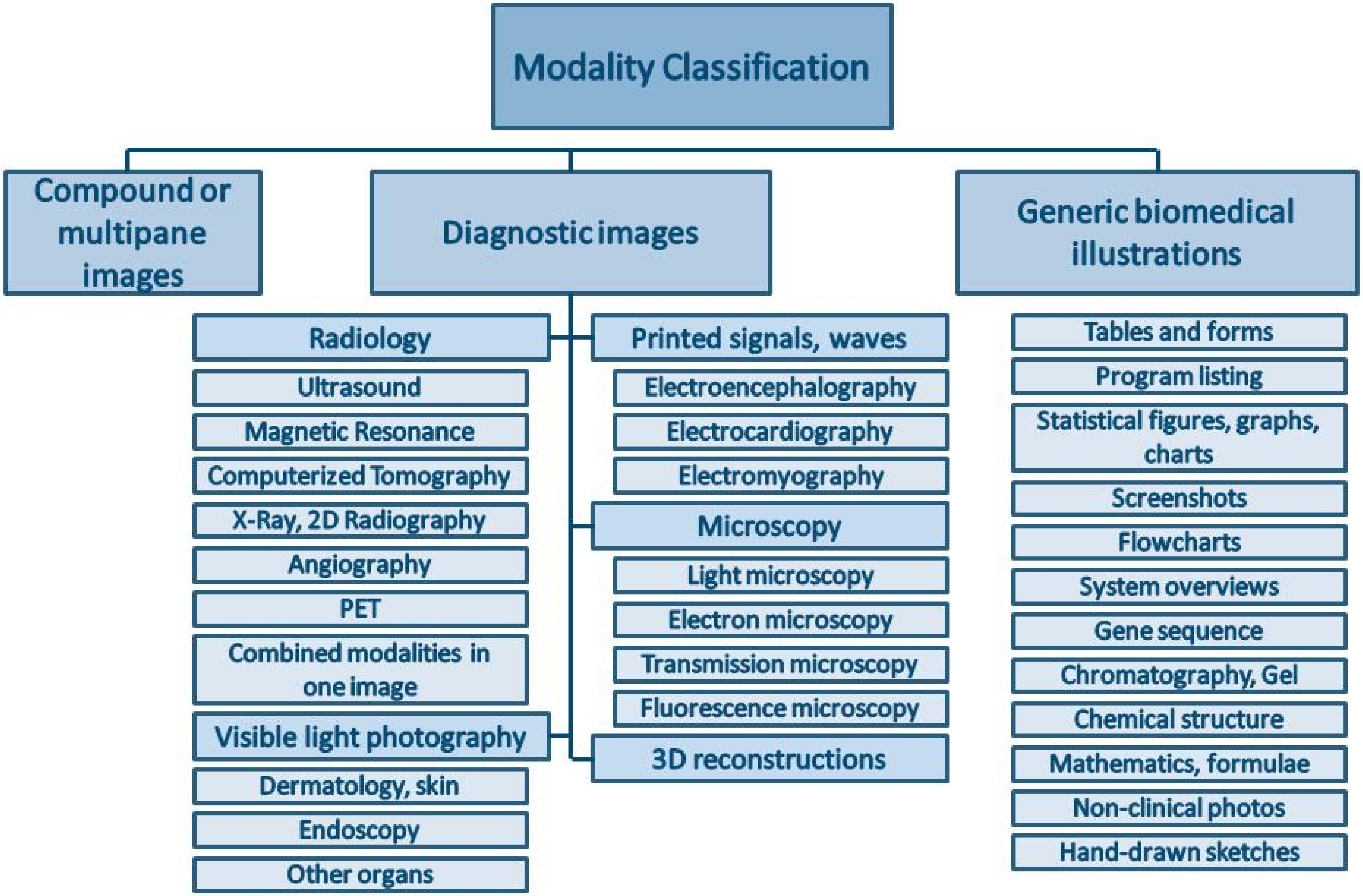
- ImageCLEF aims to provide an evaluation forum for the cross–language annotation and retrieval of images (launched in 2003)

- Statistics of this database :
  - With more than 300,000 (in .JPEG format), the total size of the database > 300 GB
  - contains PET, CT, MRI, and Ultrasound images

- Three Tasks
  - Modality classification
  - Image–based retrieval
  - Case–based retrieval

Medical Image Database available at http://www.imageclef.org/2013/medical

Modality Classification Task



- Modality Classification
  - Compound or multipane images
  - Diagnostic images
    - Radiology
      - Ultrasound
      - Magnetic Resonance
      - Computerized Tomography
      - X-Ray, 2D Radiography
      - Angiography
      - PET
      - Combined modalities in one image
    - Visible light photography
      - Dermatology, skin
      - Endoscopy
      - Other organs
    - Printed signals, waves
      - Electroencephalography
      - Electrocardiography
      - Electromyography
    - Microscopy
      - Light microscopy
      - Electron microscopy
      - Transmission microscopy
      - Fluorescence microscopy
    - 3D reconstructions
  - Generic biomedical illustrations
    - Tables and forms
    - Program listing
    - Statistical figures, graphs, charts
    - Screenshots
    - Flowcharts
    - System overviews
    - Gene sequence
    - Chromatography, Gel
    - Chemical structure
    - Mathematics, formulae
    - Non-clinical photos
    - Hand-drawn sketches

Modality is one of the most important filters that clinicians would like to be able to limit their search by.

Image based and Case-based Querying

- **Image-based retrieval** :

  This is the classic medical retrieval task.

  Similar to Query by Image Example.

  Given the query image, find the most similar images.

- **Case-based retrieval**:

  This is a more complex task; is closer to the clinical workflow.

  A case description, with patient demographics, limited symptoms and test results including imaging studies, is provided (but not the final diagnosis).

  The goal is to retrieve cases including images that might best suit the provided case description.

Challenges with Image Data

- Extracting informative features.

- Selection of relevant features.
  - Sparse methods* and dimensionality reducing techniques

- Integration of Image data with other data available
  - Early Fusion
    - Vector-based Integration
  - Intermediate Fusion
    - Multiple Kernel Learning
  - Late Fusion
    - Ensembling results from individual modalities

*Jieping Ye's SDM 2010 Tutorial on Sparse methods http://www.public.asu.edu/~jye02/Tutorial/Sparse-SDM10.pdf

Publicly Available Medical Image Repositories

Image database Name	Modalities	No. Of patients	No. Of Images	Size Of Data	Notes/Applications	Download Link
Cancer Imaging Archive Database	CT DX CR	1010	244,527	241 GB	Lesion Detection and classification, Accelerated Diagnostic Image Decision, Quantitative image assessment of drug response	https://public.cancerimagingarchive.net/ncia/dataBasketDisplay.jsf
Digital Mammography database	DX	2620	9,428	211 GB	Research in Development of Computer Algorithm to aid in screening	http://marathon.csee.usf.edu/Mammography/Database.html
Public Lung Image Database	CT	119	28,227	28 GB	Identifying Lung Cancer by Screening Images	https://eddie.via.cornell.edu/crpf.html
Image CLEF Database	PET CT MRI US	unknown	306,549	316 GB	Modality Classification , Visual Image Annotation , Scientific Multimedia Data Management	http://www.imageclef.org/2013/medical
MS Lesion Segmentation	MRI	41	145	36 GB	Develop and Compare 3D MS Lesion Segmentation Techniques	http://www.ia.unc.edu/MSseg/download.php
ADNI Database	MRI PET	2851	67,871	16GB	Define the progression of Alzheimer’s disease	http://adni.loni.ucla.edu/data-samples/acscess-data/

GENETIC DATA

Genetic Data

- The human genome is made up of DNA which consists of four different chemical building blocks (called bases and abbreviated A, T, C, and G).

- It contains 3 billion pairs of bases and the particular order of As, Ts, Cs, and Gs is extremely important.

- Size of a single human genome is about 3GB.

- Thanks to the Human Genome Project (1990-2003)
  - The goal was to determine the complete sequence of the 3 billion DNA subunits (bases).
  - The total cost was around $3 billion.

Genetic Data

- The whole genome sequencing data is currently being annotated and not many analytics have been applied so far since the data is relatively new.

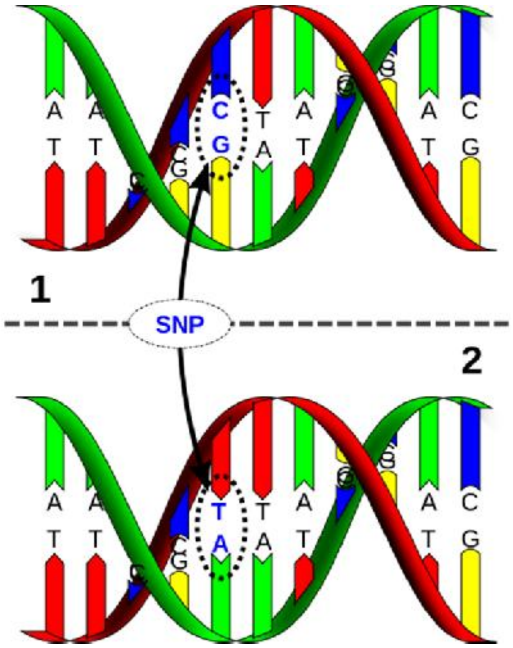
- Several publicly available genome repositories. http://aws.amazon.com/1000genomes/

- It costs around $5000 to get a complete genome. It is still in the research phase. Heavily used in the cancer biology.

- In this tutorial, we will focus on Genome-Wide Association Studies (GWAS).
  - It is more relevant to healthcare practice. Some clinical trials have already started using GWAS.
  - Most of the computing literature (in terms of analytics) is available for the GWAS. It is still in rudimentary stage for whole genome sequences.

Genome-Wide Association Studies (GWAS)

- Genome-wide association studies (GWAS) are used to identify common genetic factors that influence health and disease.
- These studies normally compare the DNA of two groups of participants: people with the disease (cases) and similar people without (controls). (One million Loci)
- Single nucleotide polymorphisms (SNPs) are DNA sequence variations that occur when a single nucleotide (A,T,C,or G) in the genome sequence differs between individuals.
- SNPs occur every 100 to 300 bases along the 3-billion-base human genome.



Epistasis Modeling

- For simple Mendelian diseases, single SNPs can explain phenotype very well.

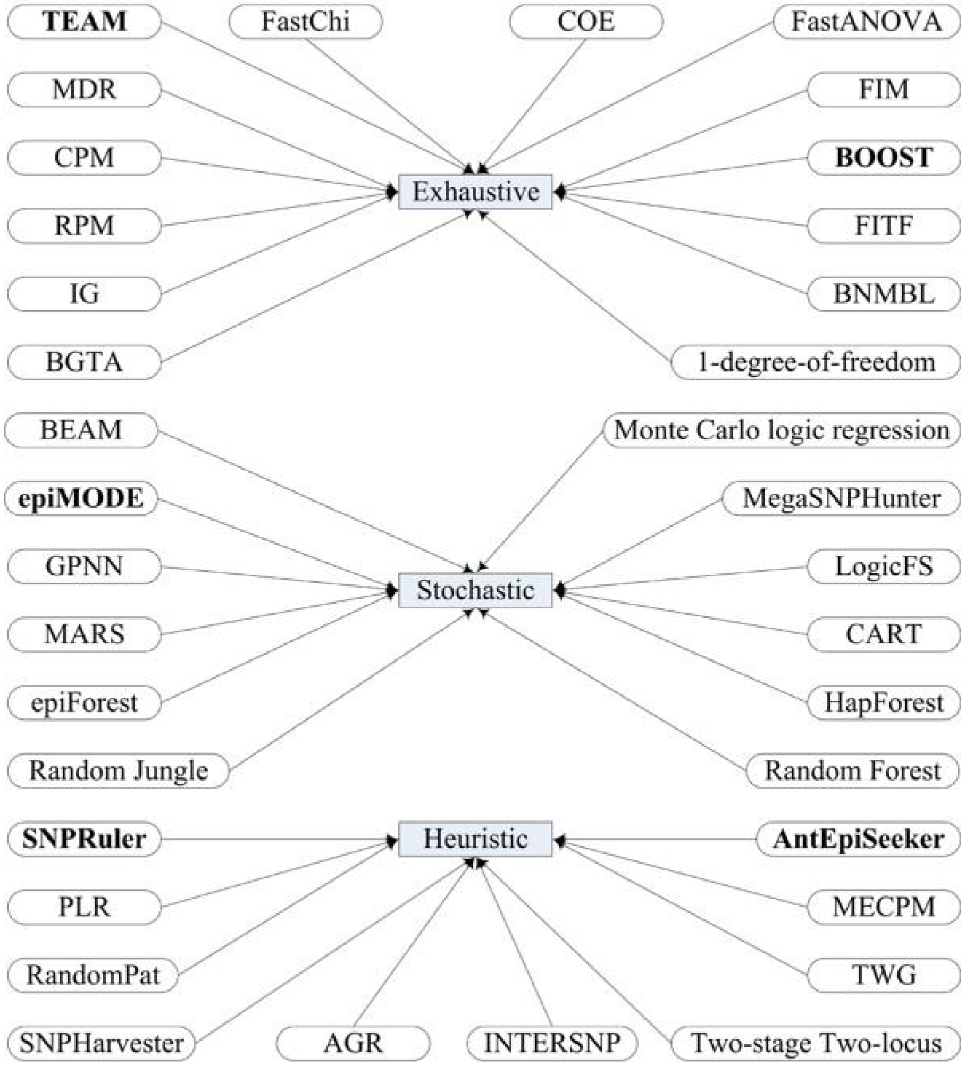
- The complex relationship between genotype and phenotype is inadequately described by marginal effects of individual SNPs.

- Increasing empirical evidence suggests that interactions among loci contribute broadly to complex traits.

- The difficulty in the problem of detecting SNP pair interactions is the **heavy computational burden.**
  - To detect pairwise interactions from 500,000 SNPs genotyped in thousands of samples, a total of 1.25 X 10^{11}statistical tests are needed.

Epistasis Detection Methods

- Exhaustive
  - Enumerates all *K*-locus interactions among SNPs.
  - Efficient implementations mostly aiming at reducing computations by eliminating unnecessary calculations.
- Non-Exhaustive
  - Stochastic: randomized search. Performance lowers when the # SNPs increase.
  - Heuristic: greedy methods that do not guarantee optimal solution.



Shang, Junliang, et al. "Performance analysis of novel methods for detecting epistasis." *BMC bioinformatics* 12.1 (2011): 475.

Sparse Methods for SNP data analysis

- Successful identification of SNPs strongly predictive of disease promises a better understanding of the biological mechanisms underlying the disease.
- Sparse linear methods have been used to fit the genotype data and obtain a selected set of SNPs.
- Minimizing the squared loss function (*L*) of *N* individuals and *p* variables (SNPs) is used for linear regression and is defined as

$$L(\ {}_0,\ ) = \frac{1}{2}\sum_{i=1}^{N}(y_i - \ {}_0 - x_i^T\ )^2 + \text{a}\sum_{j=1}^{p}\left|\ {}_j\right|$$

where $x_i \in \mathbb{R}^p$ are inputs for the i^{th} sample, $y \in \mathbb{R}^N$ is the N vector of outputs, $\beta_0 \in \mathbb{R}$ is the intercept, $\beta \in \mathbb{R}^p$ is a p-vector of model weights, and λ is user penalty.

- Efficient implementations that scale to genome-wide data are available.
- SparSNP package http://bioinformatics.research.nicta.com.au/software/sparsnp/

Wu, Tong Tong, et al. "Genome-wide association analysis by lasso penalized logistic regression." *Bioinformatics* 25.6 (2009): 714-721.

Public Resources for Genetic (SNP) Data

- The Wellcome Trust Case Control Consortium (WTCCC) is a group of 50 research groups across the UK which was established in 2005.
- Available at http://www.wtccc.org.uk/
- Seven different diseases: bipolar disorder (1868 individuals), coronary heart disease (1926 individuals), Crohn’s disease (1748 individuals), hypertension (1952 individuals), rheumatoid arthritis (1860 individuals), type I diabetes (1963 individuals) or type II diabetes (1924 individuals).
- Around 3,000 healthy controls common for these disorders. The individuals were genotyped using Affymetrix chip and obtained approximately 500K SNPs.

- The database of Genotypes and Phenotypes (dbGaP) maintained by National Center of Biotechnology Information (NCBT) at NIH.
- Available at http://www.ncbi.nlm.nih.gov/gap

BEHAVIORAL AND PUBLIC HEALTH DATA

Epidemiology Data

- The Surveillance Epidemiology and End Results Program (SEER) at NIH.
- Publishes cancer incidence and survival data from population-based cancer registries covering approximately 28% of the population of the US.
- Collected over the past 40 years (starting from January 1973 until now).
- Contains a total of 7.7M cases and >350,000 cases are added each year.
- Collect data on patient demographics, tumor site, tumor morphology and stage at diagnosis, first course of treatment, and follow-up for vital status.

Usage:

- Widely used for understanding disparities related to race, age, and gender.
- Can be used to overlay information with other sources of data (such as water/air pollution, climate, socio-economic) to identify any correlations.
- Can not be used for predictive analysis, but mostly used for studying trends.

SEER database is available at http://seer.cancer.gov/

Social Media can Sense Public Health !!

During infectious disease outbreaks, data collected through health institutions and official reporting structures may not be available for weeks, hindering early epidemiologic assessment. Social media can get it in near real-time.

Twitter messaging correlated with cholera outbreak

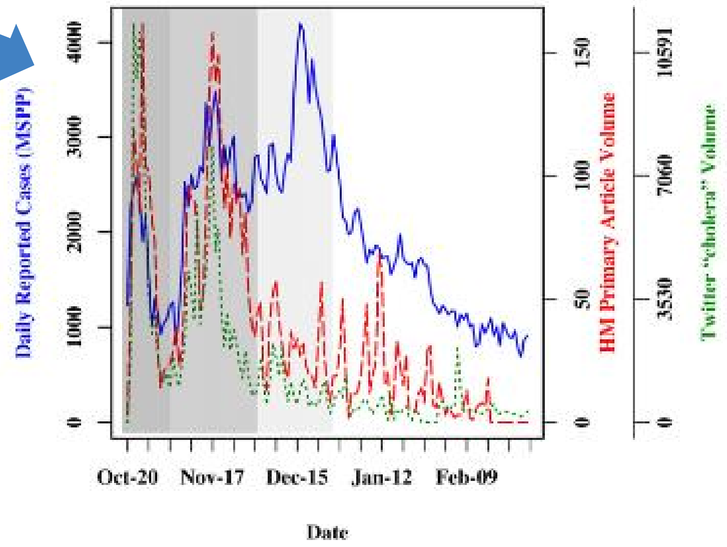
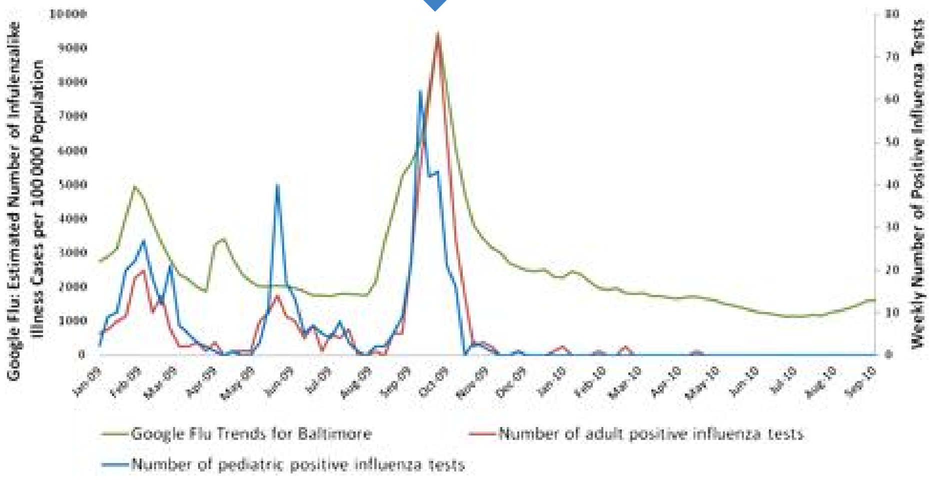


FIGURE 2. Daily reported case data for all departments from the Haiti Ministry of Health (solid), daily volume of primary HealthMap alerts (dashed), and daily volume of Twitter posts containing the word “cholera” or “#cholera” (dotted). Each curve has an initial peak at the onset of the outbreak (dark grey), and a peak during the time that Hurricane Tomas affected Haiti (medium grey). The first 100 days of the outbreak are shaded in light grey. Ministère de la Santé Publique et de la Population (MSPP) case counts peak again in late December, although HealthMap and Twitter volume only have daily variations during this time.

Google Flu Trends correlated with Influenza outbreak



Dugas, Andrea Freyer, et al. "Google Flu Trends: correlation with emergency department influenza rates and crowding metrics." *Clinical infectious diseases* 54.4 (2012): 463-469.

Chunara, Rumi, Jason R. Andrews, and John S. Brownstein. "Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak." *American Journal of Tropical Medicine and Hygiene* 86.1 (2012): 39.

Social Networks for Patients

- **PatientsLikeMe**[1] is a patient network is an online data sharing platform started in 2006; now has more than 200,000 patients and is tracking 1,500 diseases.

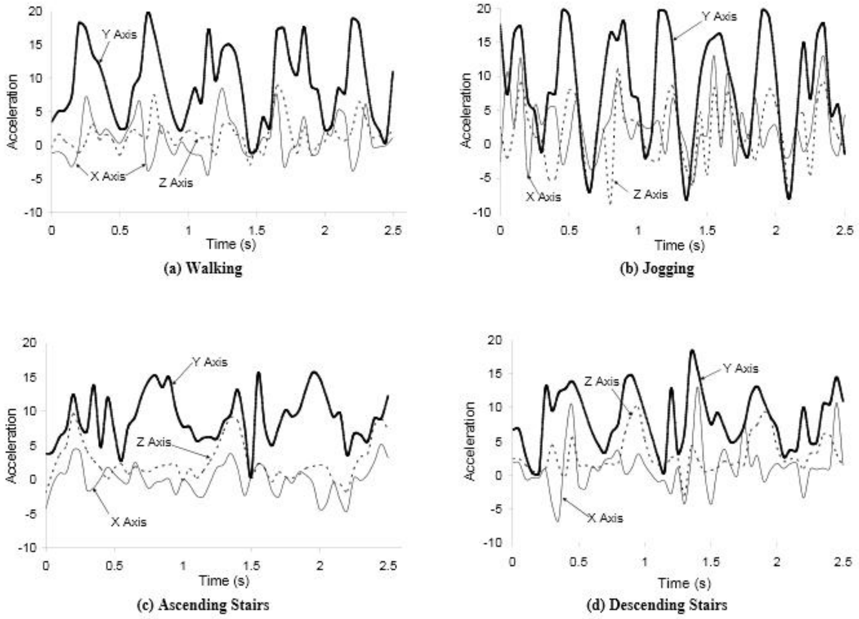
OBJECTIVE: “Given my status, what is the best outcome I can hope to achieve, and how do I get there?”

- People connect with others who have the same disease or condition, track and share their own experiences, see what treatments have helped other patients like them, gain insights and identify any patterns.
- Patient provides the data on their conditions, treatment history, side effects, hospitalizations, symptoms, disease-specific functional scores, weight, mood, quality of life and more on an ongoing basis.
- Gaining access to the patients for future clinical trials.

[1] http://www.patientslikeme.com/

Home Monitoring and Sensing Technologies

- Advancements in sensing technology are critical for developing effective and efficient home-monitoring systems
- Sensing devices can provide several types of data in real-time.
- Activity Recognition using Cell Phone Accelerometers



(a) Walking

(b) Jogging

(c) Ascending Stairs

(d) Descending Stairs

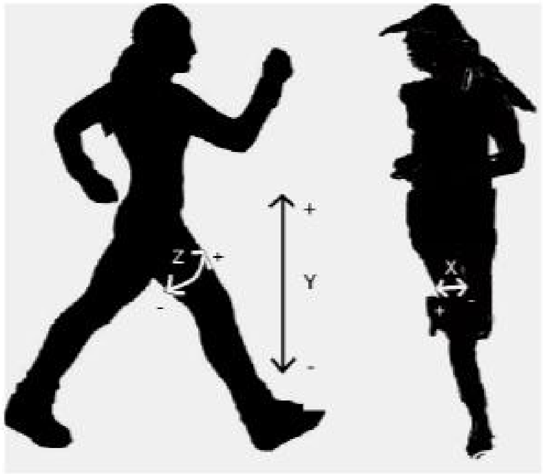


Figure 1: Axes of Motion Relative to User

Kwapisz, Jennifer R., Gary M. Weiss, and Samuel A. Moore. "Activity recognition using cell phone accelerometers." *ACM SIGKDD Explorations Newsletter* 12.2 (2011): 74-82.
Rashidi, Parisa, et al. "Discovering activities to recognize and track in a smart environment." *Knowledge and Data Engineering, IEEE Transactions on* 23.4 (2011): 527-539.

Public Health and Behavior Data Repositories

Dataset	Link	Description
Behavioral Risk Factor Surveillance System (BRFSS)	http://www.cdc.gov/brfss/technical_infodata/index.htm	Healthcare survey data: smoking, alcohol, lifestyle (diet, exercise), major diseases (diabetes, cancer), mental illness
Ohio Hospital Inpatient/Outpatient Data	http://publicapps.odh.ohio.gov/pwh/PWHMain.aspx?q=021813114232	Hospital: number of discharges, transfers, length of stay, admissions, transfers, number of patients with specific procedure codes
US Mortality Data	http://www.cdc.gov/nchs/data_access/cmf.htm	Mortality information on county-level
Human Mortality Database	http://www.mortality.org/	Birth, death, population size by country
Utah Public Health Database	http://ibis.health.utah.gov/query	Summary statistics for mortality, charges, discharges, length of stay on a county-level basis
Dartmouth Atlas of Health Care	http://www.dartmouthatlas.org/tools/downloads.aspx	Post discharge events, chronically ill care, surgical discharge rate

Thanks to Prof. Joydeep Ghosh from UT Austin for providing this information.

CONCLUDING REMARKS

Final Thoughts

Big data could save the health care industry up to $450 billion, but other things are important too.

- **Right living:** Patients should take more active steps to improve their health.
- **Right care:** Developing a coordinated approach to care in which all caregivers have access to the same information.
- **Right provider:** Any professionals who treat patients must have strong performance records and be capable of achieving the best outcomes.
- **Right value:** Improving value while simultaneously improving care quality.
- **Right innovation:** Identifying new approaches to health-care delivery.

"Stakeholders will only benefit from big data if they take a more holistic, patient-centered approach to value, one that focuses equally on health-care spending and treatment outcomes,"

McKinsey report available at: http://www.mckinsey.com/insights/health_systems/the_big-data_revolution_in_us_health_care

Conclusion

- Big data analytics is a promising right direction which is in its infancy for the healthcare domain.

- Healthcare is a data-rich domain. As more and more data is being collected, there will be increasing demand for big data analytics.

- Unraveling the “Big Data” related complexities can provide many insights about making the right decisions at the right time for the patients.

- Efficiently utilizing the colossal healthcare data repositories can yield some immediate returns in terms of patient outcomes and lowering care costs.

- Data with more complexities keep evolving in healthcare thus leading to more opportunities for big data analytics.

Acknowledgements

- Funding Sources
  - – National Science foundation
  - – National Institutes of Health
  - – Susan G. Komen for the Cure
  - – Delphinus Medical Technologies
  - – IBM Research

Thank You

Questions and Comments

Feel free to email questions or suggestions to

jimeng@cs.cmu.edu

reddy@cs.wayne.edu