Natural Language Processing for Clinical Informatics and Translational Research Informatics

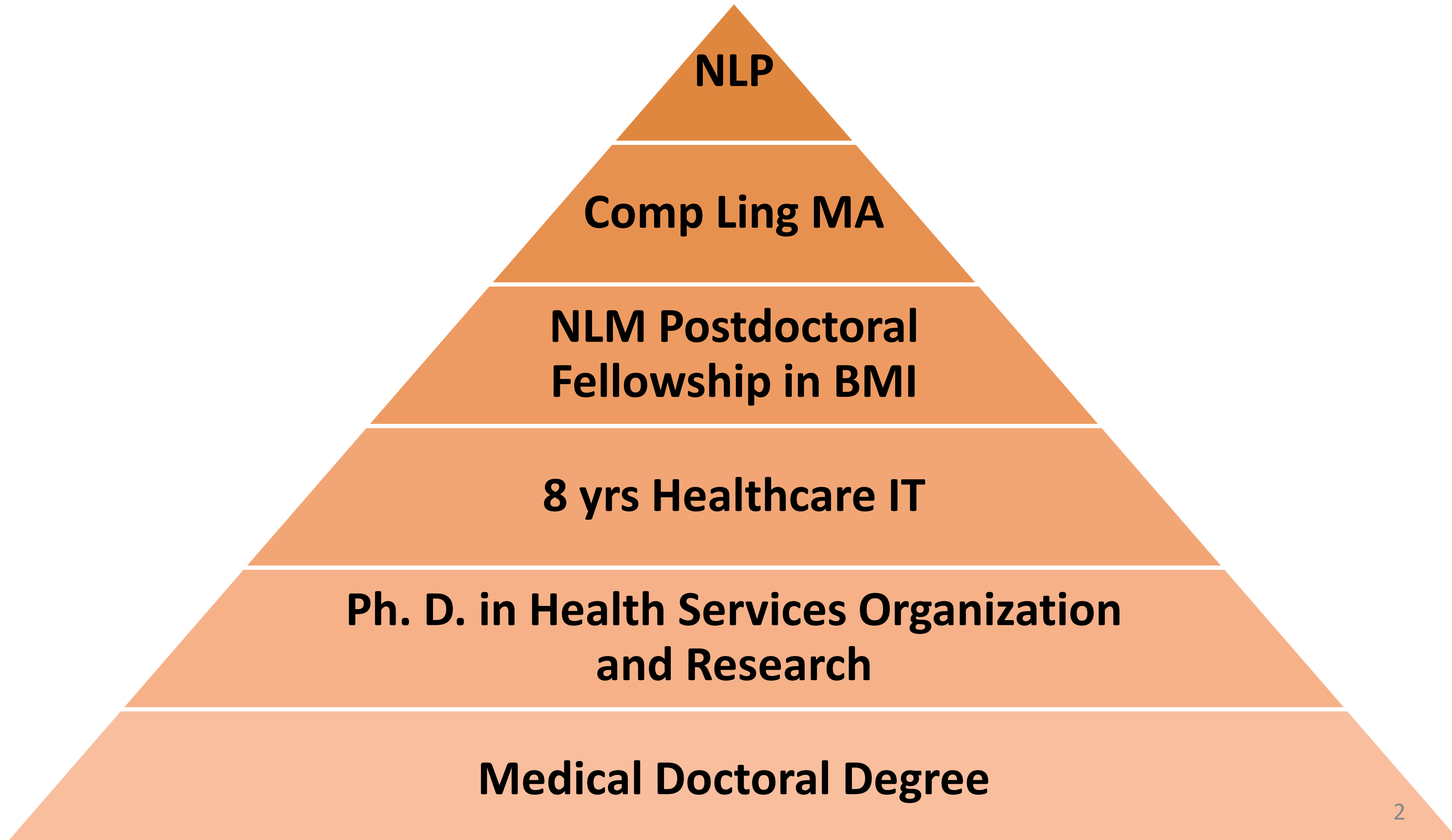
Imre Solti, M. D., Ph. D.

solti@uw.edu

K99 Fellow in Biomedical Informatics

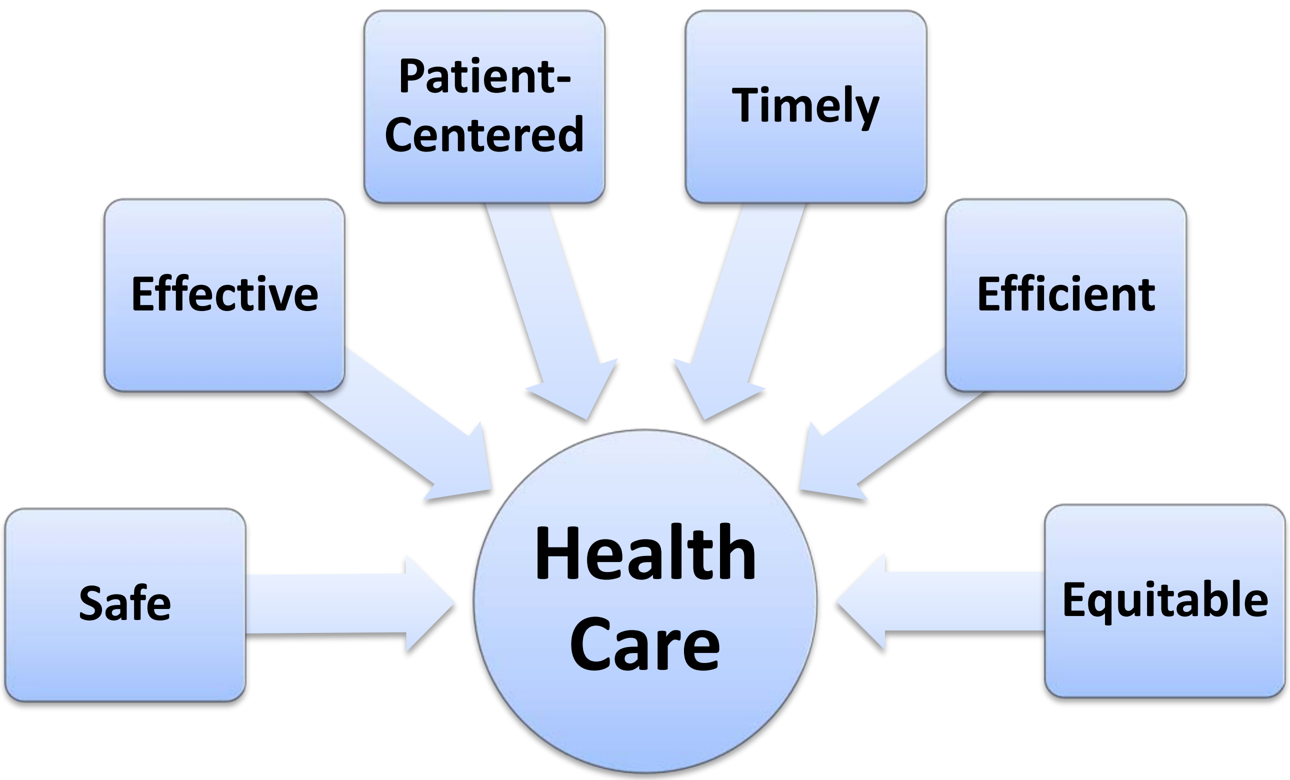
University of Washington

Background for Clinical Natural Language Processing (NLP)



Career Interest

NLP as strategic tool to achieve the six aims of the Institute of Medicine



Research Interests

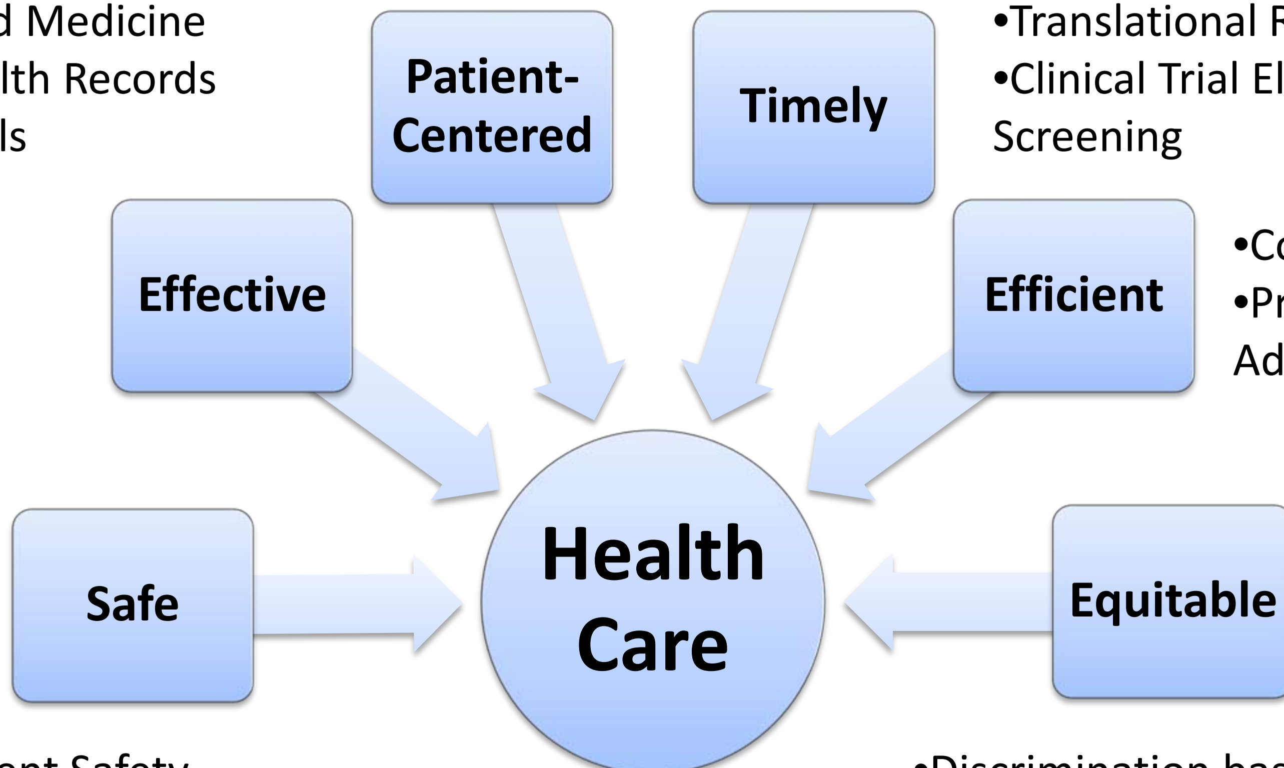
NLP for Clinical Informatics and Translational Research Informatics

- Individualized Medicine
- Personal Health Records
- Patient Portals

- Translational Research
- Clinical Trial Eligibility Screening

- Comparative-Effectiveness

- Cost of Care
- Present On Admission



- Patient Safety
- Adverse Drug Events
- Pharmacovigilance

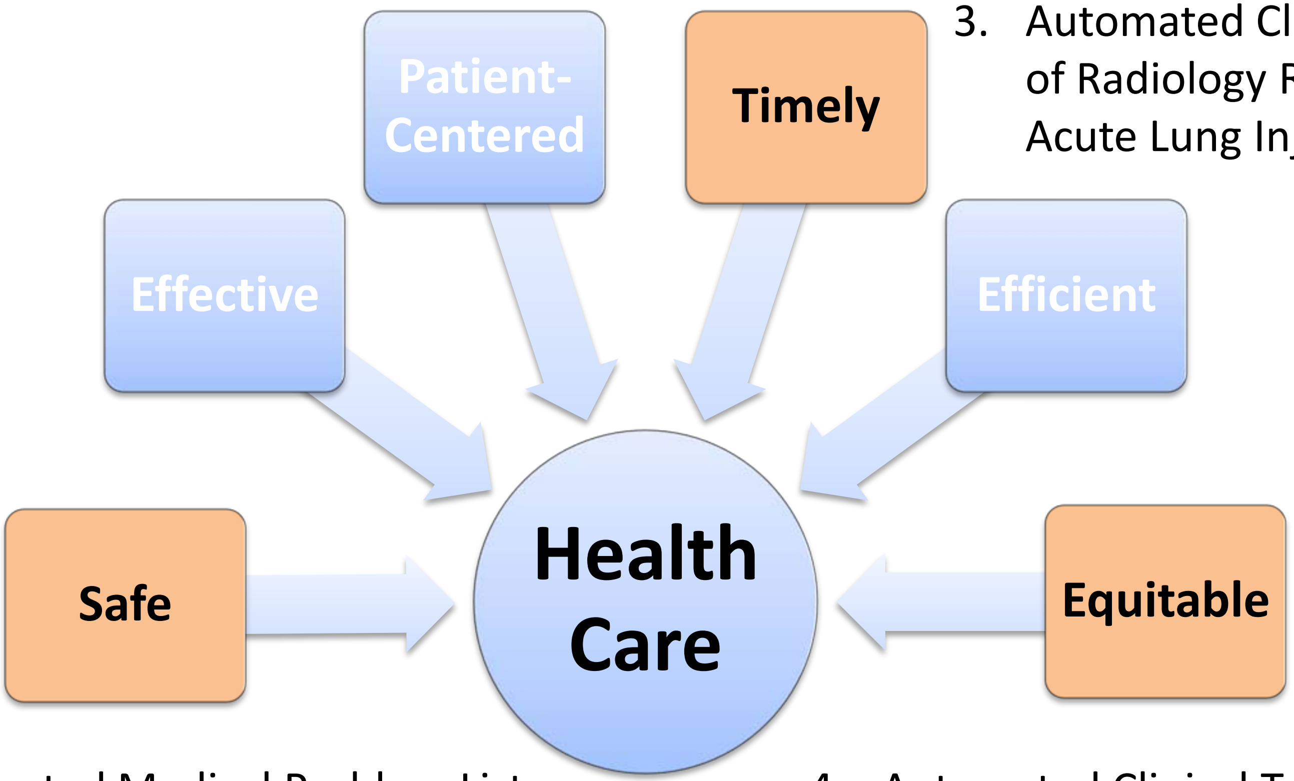
- Discrimination based on Age or Ethnicity

Research Interests - Summary

- **Information Extraction** from unstructured clinical text -> Linking phenotype and genotype
- **Document Classification**
- **Data Mining**

Use Cases for Today's Presentation

NLP Research Use Cases for the Electronic Medical Record



3. Automated Classification of Radiology Reports for Acute Lung Injury

1. Semi-Automated Medical Problem List
2. Extraction of Medication Information

4. Automated Clinical Trial Eligibility Screening

Collaborators

- University of Washington
  - **Eithon Cadag**, Ph. D. – Biomedical Informatics
  - **John Gennari**, Ph. D. - Biomedical Informatics
  - **Scott Halgrim**, M. A. – Computational Linguistics
  - **Tom Payne**, M. D. – IT Services Medical Center
  - **Peter Tarczy-Hornoch**, M. D. - Biomedical Informatics
  - **Mark Wurfel**, M. D. – Pulmonary and Critical Care Med
  - **Fei Xia**, Ph. D. – Computational Linguistics

- External Investigators
  - **University of Pittsburgh**
  - **Columbia**
  - Albany/MIT, **i2b2** (Informatics for Integrating Biology and the Bedside)

Definitions[1]

- Natural Language Processing (NLP):

  NLP research focuses on building computational models for understanding natural (human) language.
- Information Extraction (IE):

  IE involves extracting predefined types of information from text. Subfield of NLP.
- Named Entity Recognition (NER):

  Recognizing expressions denoting entities (i.e., Named Entities), such as diseases in free text documents. Subfield of IE.
- Information Retrieval (IR):

  Information retrieval (IR) is focused on finding documents.

[1] Meystre, S. M., et al., "Extracting information from textual documents in the electronic health record: a review of recent research." Yearb Med Inform. 2008:128-44.

Definitions[1] – Cont.

- Document Classification:

  Assigning electronic documents to one or more categories.
- Biomedical Text:

  Text that appears in books, articles, literature abstracts.
- Clinical Text:

  Texts written by clinicians in the clinical setting.
- Biomedical-NLP:

  NLP for biomedical text.
- Clinical-NLP:

  NLP for the clinical text.

[1] Meystre, S. M., et al., "Extracting information from textual documents in the electronic health record: a review of recent research." Yearb Med Inform. 2008:128-44.

Agenda for Today

Past Projects:

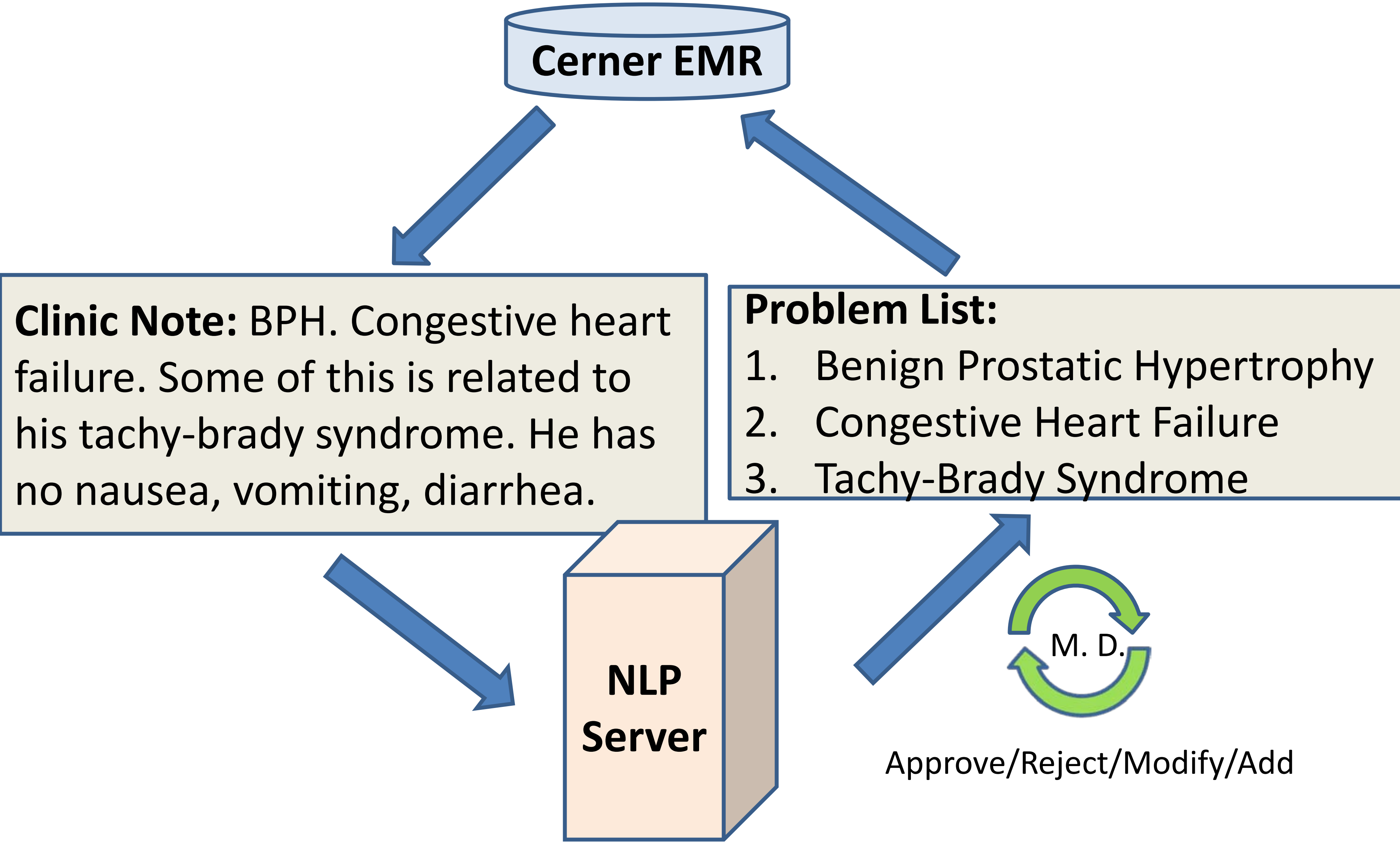
1. Semi-Automated Medical Problem List: Clinical-NLP, IE, NER - 1 Slide
2. Extraction of Medication Information: Clinical-NLP, IE, NER - 1 Slide
3. **Classification of Radiology Reports for Acute Lung Injury**: Clinical Document Classification

Future Project:

4. ***Automated Clinical Trial Eligibility Screening**: Clinical NLP, Biomedical-NLP, IE, NER, Document Classification

***Grant funded**

Semi-Automated Medical Problem List



Solti I., et al. “Building an automated problem list based on natural language processing: lessons learned in the early phase of development. “AMIA Annu Symp Proc. 2008 Nov 6:687-91.

Automated Extraction of Medication Information

DISCHARGE MEDICATIONS:
...
Additionally, **A spirin** **1-2 tablets** **p.o.** **q 4 prn** during next **ten days** for his **pain**, ...

IE

- Medication: **"Aspirin"**
- Dose: **"1-2 tablets"**
- Mode of Administration: **"p. o."**
- Frequency: **"q 4 prn"**
- Duration: **"ten days"**
- Medical Reason: **"pain"**

MaxEnt Algorithm

ML Features

Hybrid IE System

Handcrafted Rules

RxNorm Lexicon

UW – i2b2 Collaboration

i2b2 Challenge

Building System

Challenge

Community Annotation

MA Thesis

Classification of Radiology Reports for Acute Lung Injury (ALI)

Motivation

- 30 % Mortality
- Delayed manual chest x-ray classification

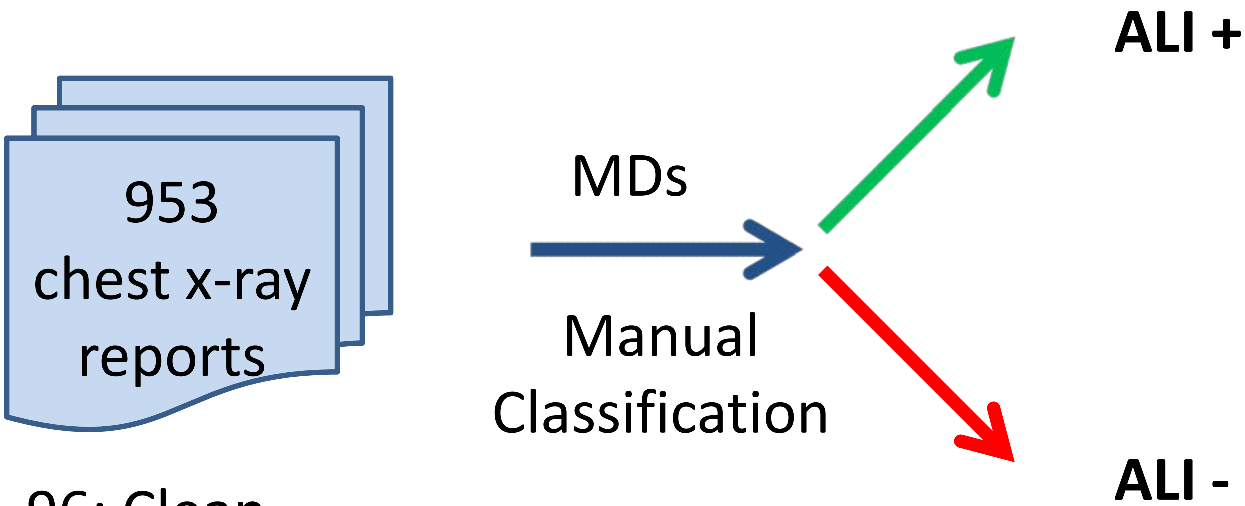
Aims

- Build NLP-based classifier
- Intuitive link: Machine Learning – Clinical Expertise

Methods

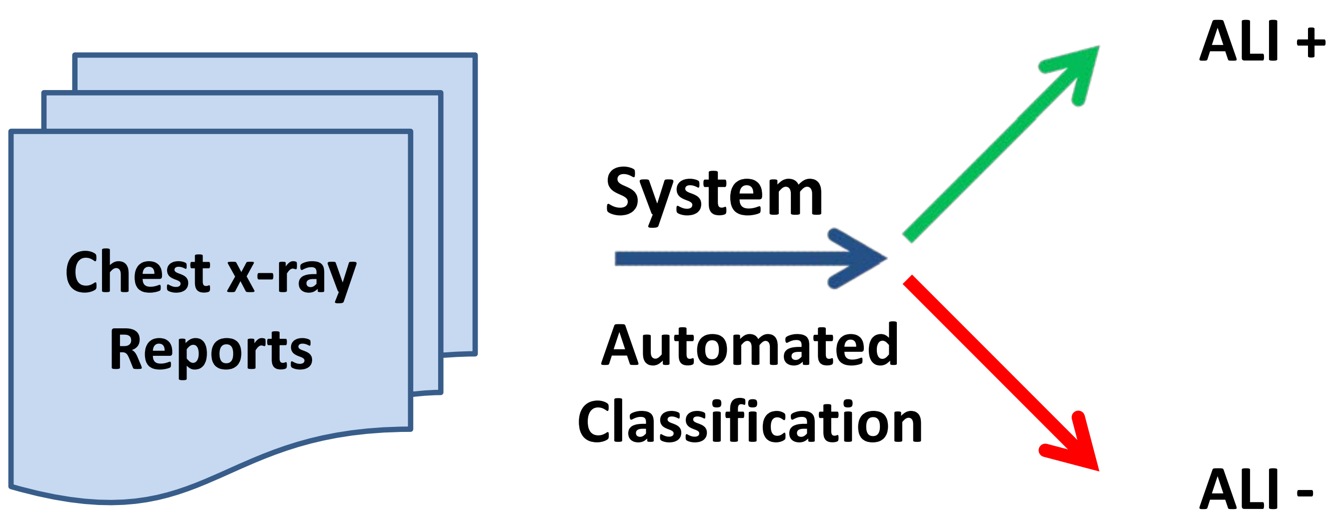
- Keywords
- Maximum Entropy: Character n-grams

Data (Corpus) and Gold Standard



96: Clean
857: Noisy

Task for Automated ALI Classification



Sample Report

Tubes and lines: satisfactory position and alignment

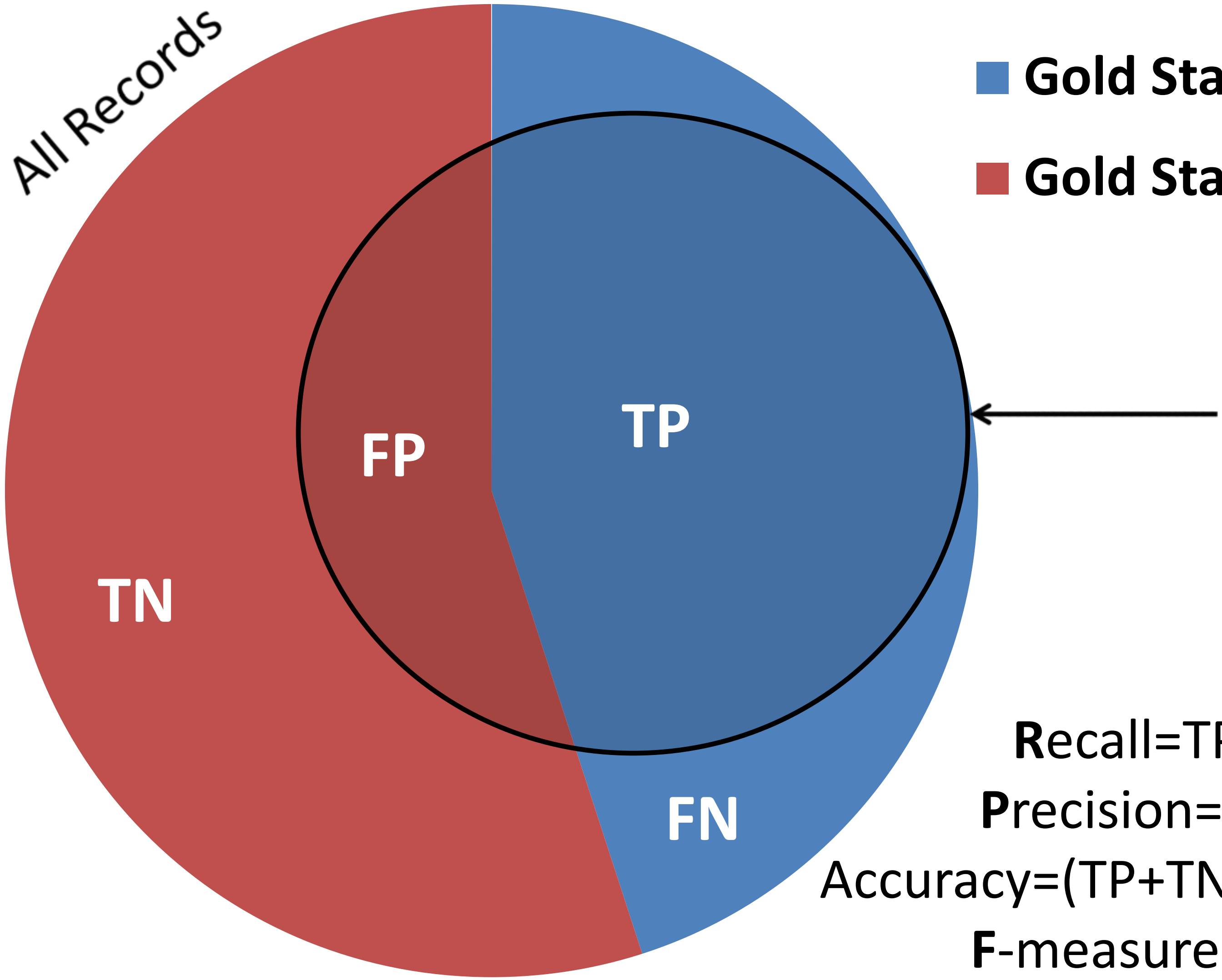
Lungs: The lung volumes are low and unchanged. There are diffuse, bilateral opacities that are worsened.

Pneumothorax: none

Effusions: none

Measurement Metrics

Recall, Precision, F-measure, Accuracy



Recall=TP/(TP+FN)

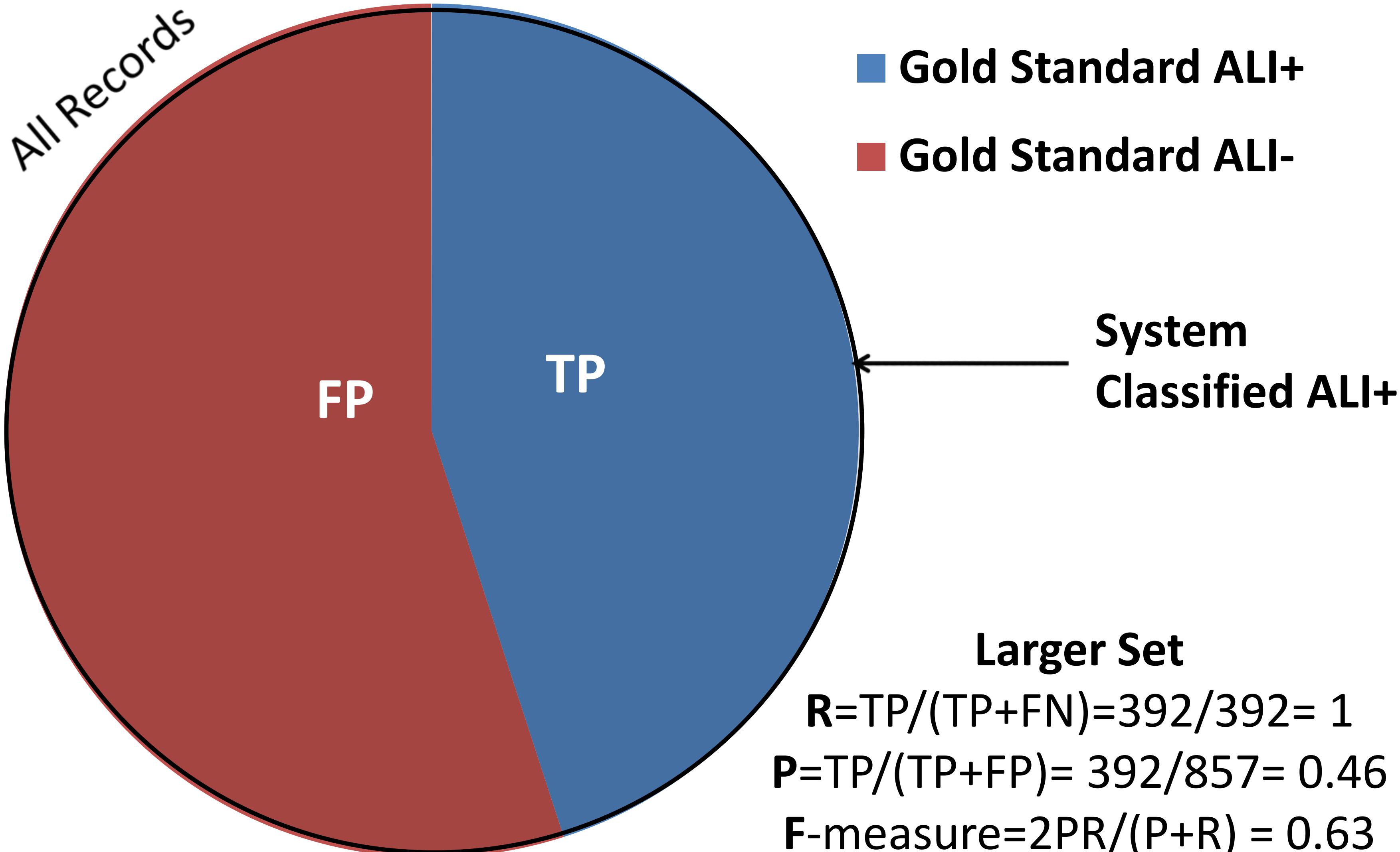
Precision=TP/(TP+FP)

Accuracy=(TP+TN)/(TP+TN+FP+FN)

F-measure=2PR/(P+R)

Baseline

No Processing: Assign ALI+ to Every Report



Larger Set

R=TP/(TP+FN)=392/392= 1

P=TP/(TP+FP)= 392/857= 0.46

F-measure=2PR/(P+R) = 0.63

Gold Standard – Smaller Corpus

Reviewer	R	P	F
1	0.94	0.98	0.96
2	0.98	0.91	0.94
3	0.80	0.95	0.87
4	0.80	0.98	0.88
5	0.62	1.00	0.77
6	1.00	0.83	0.91
7	0.92	0.94	0.93
8	0.70	0.92	0.80
9	0.70	1.00	0.82
10	0.96	0.96	0.96
11	0.92	0.98	0.95

List of Keywords (Sample)

Phrase	Weight/3	Weight/10
edema	2.5	8
lung opacities	2	5.5
diffuse	3	10
bilateral	3	10

48 Key Phrases

Keyword & Weight-Based Results

Method	R	P	F	Acc
96-raw	0.88	0.83	0.85	0.844
96-w3	0.82	0.85	0.84	0.833
96-w10	0.72	0.88	0.80	0.800
Baseline	1	0.46	0.63	0.46

MaxEnt Character n-gram Features

- Unigram, Bigram, … 6-gram
- “diffuse”, 6-gram, sliding window

nnnn_d

nnn_di

nn_dif

n_diff

_diffu

diffus

iffuse

ffuse_

fuse_n

etc …

MaxEnt Results (Smaller Corpus)

System	R	P	F	Acc
word	0.83	0.78	0.80	**0.81**
n1	0.62	0.58	0.60	0.63
n2	0.67	0.81	0.73	0.76
n3	0.82	0.85	**0.84**	**0.82**
n4	0.85	0.97	**0.91**	**0.88**
n5	0.77	0.73	0.75	0.78
n6	0.84	0.82	**0.83**	**0.85**
Baseline	1	0.46	0.63	0.46

MaxEnt vs Keyword

System	R	P	F	Acc
Raw	**0.88**	0.83	**0.85**	0.84
W3	0.82	**0.85**	0.84	0.83
n3	0.82	**0.85**	0.84	0.82
n4	**0.85**	**0.97**	**0.91**	**0.88**
n6	0.84	0.82	0.83	**0.85**
Baseline	1	0.46	0.63	0.46

ROC statistics - Not significant difference Keyword vs MaxEnt

Visualizing Machine Learning Features - MaxEnt

Present on the 48-Phrase List	
N-gram Feature	**Clinical Phrase**
edema_	edema
a_and_	edem**a and**
ffuse_	di**ffuse**
teral_	bila**teral**
y_opac	patch**y opac**ities
al_opa	bilater**al opa**cities
Missing from 48-Phrase List	
perihi	**perihi**lar

Limitations

1. Two corpora (Selection and GS Criteria)
2. Not tested – Other ALI Research Team Corpora
3. Features limited to n-grams
4. Different performance peaks (96 vs 857-set)

Related Work ALI Classification

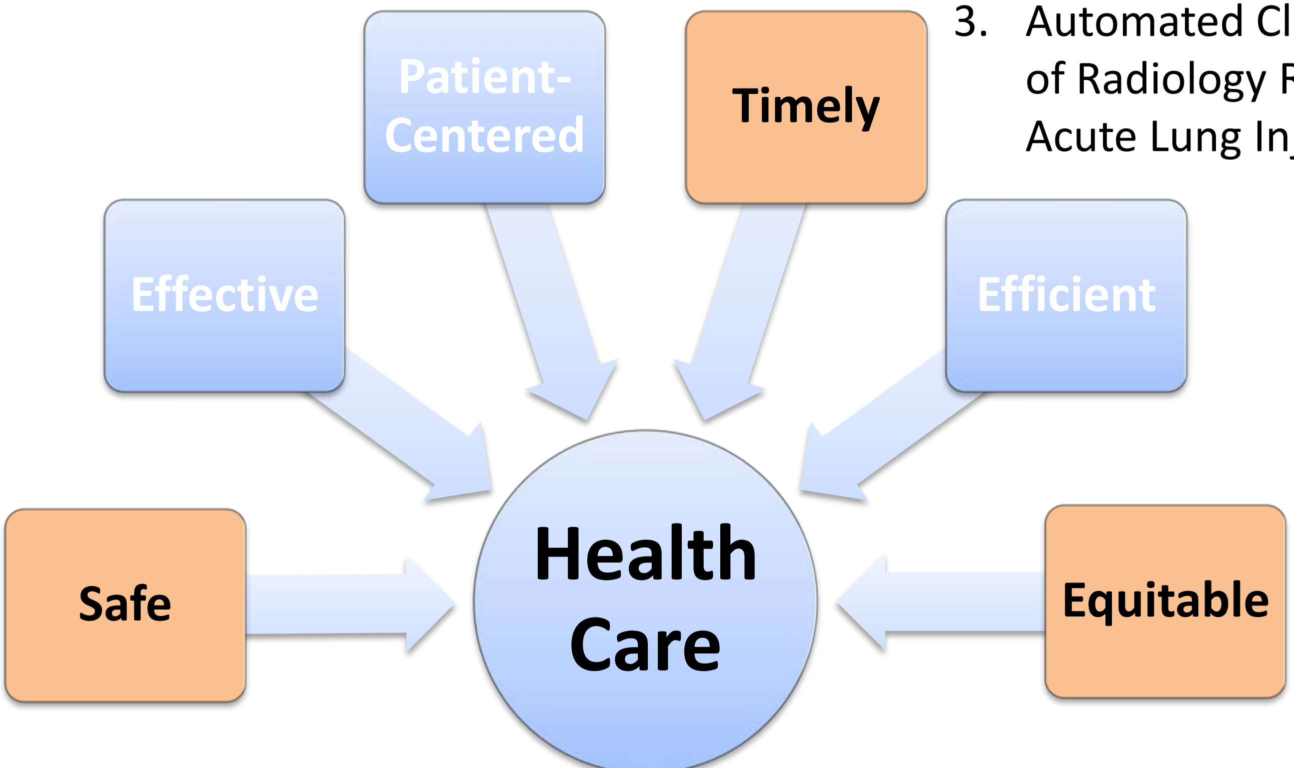
- Herasevich et al., - Mayo Clinic, Rochester (2009)
- Azzam et al., - UPenn (2009)
- Rule-based systems, focus -> ALI screening not on NLP component
- No details -> Not directly comparable

Conclusions for ALI Classification

1. **Aims achieved:**
   I. Built NLP-based classifier(s)
   II. Visualized ML features for clinicians
2. **Advantages and disadvantages:** Keyword and ML-based systems ->
3. **What approach is better?**

Use Cases for Today's Presentation

NLP Research Use Cases for the Electronic Medical Record

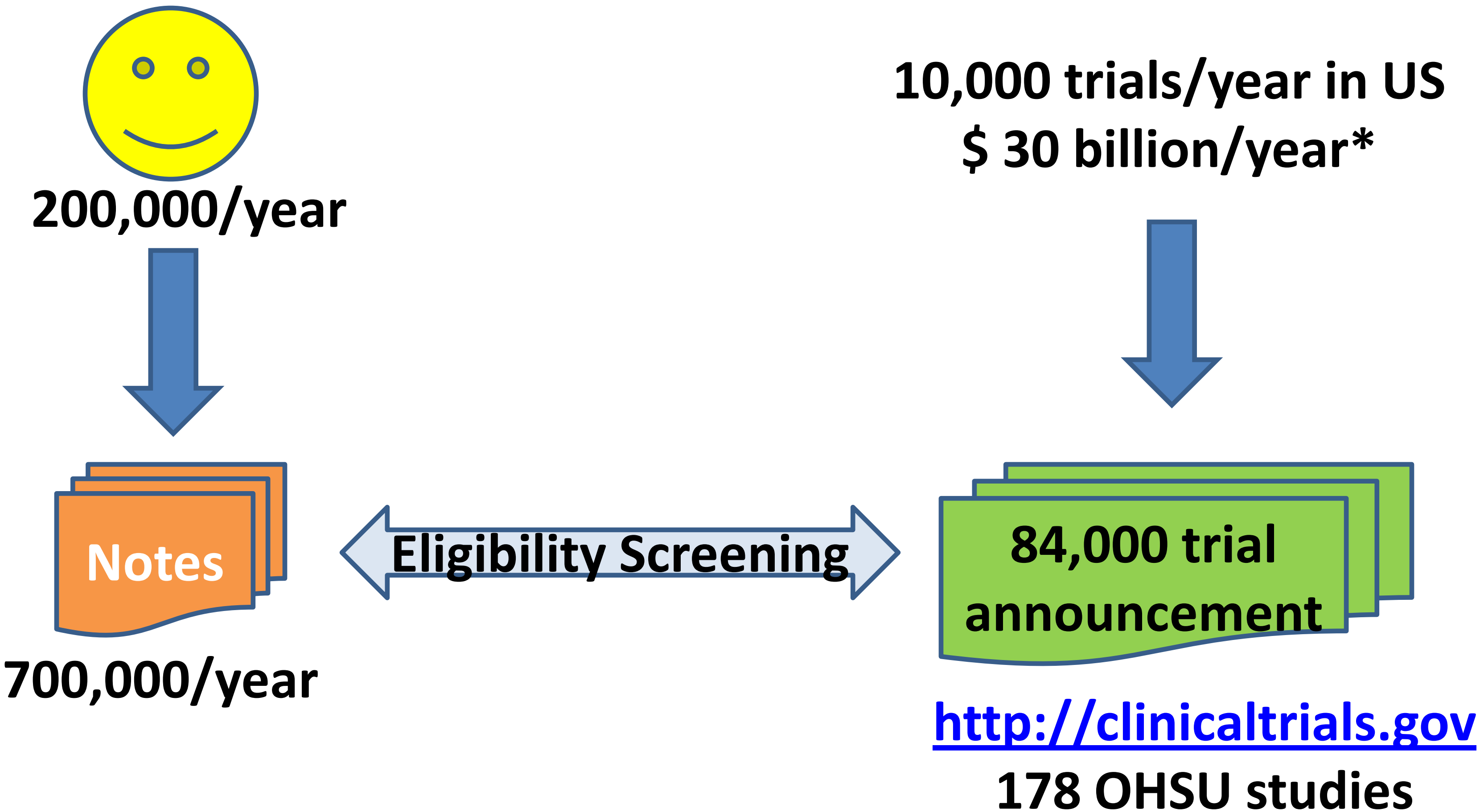


3. Automated Classification of Radiology Reports for Acute Lung Injury

1. Semi-Automated Medical Problem List
2. Extraction of Medication Information

4. Automated Clinical Trial Eligibility Screening

Automated Clinical Trial Eligibility Screening - Task



*Editorial - The Cost of Clinical Trials. *Drug Discovery & Development* . March 01, 2007

Automated Patient-Centered Clinical Trial Eligibility Screening

Background and Significance:

- **Low Rate:** 4% adult cancer patients
- **Physician Bias:** older age, minority status
- **Not Mentioned:** 25% b cc surg -> 0 offer,
  40% -> 1-10% offer

Aims:

- Identify concept elements
- Build inf application to extract and match
- Interactive input module
- Evaluation of performance

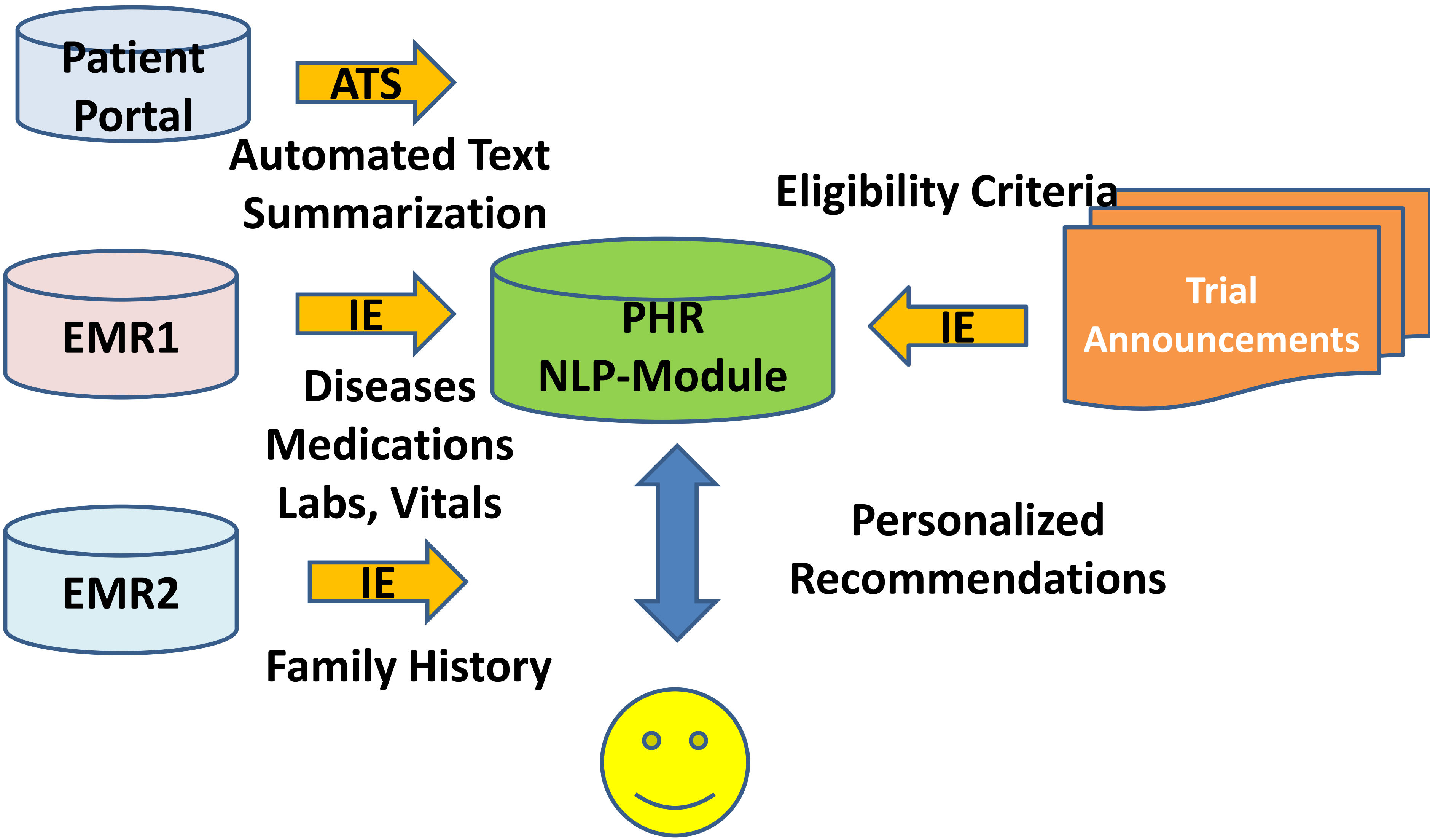
Related Work

- Protocol Authoring Tools
- Standardized Terminology
  - **C**linical **D**ata **I**nterchange **S**tandards **C**onsortium
  - **B**iomedical **R**esearch **I**ntegrated **D**omain **G**roup
  - HL7
  - Trial Bank/Open Trial Bank – Ida Sim
  - Columbia – Patel and Weng
- Cincinnati - Embi
- Others…

Excerpts – Trial Announcement

- **First-degree relative with** bilateral breast cancer who developed the **first breast cancer at ≤ 50 years of age**
- **Postmenopausal**, defined as **at least 1 of the following:**
  - **Over 60 years of age**
  - **Bilateral oophorectomy**
  - **≤ 60 years of age with** a uterus and **amenorrhea for at least 12 months**
- **No cancer within** the **past 5 years except nonmelanoma skin cancer or carcinoma in situ of the cervix**

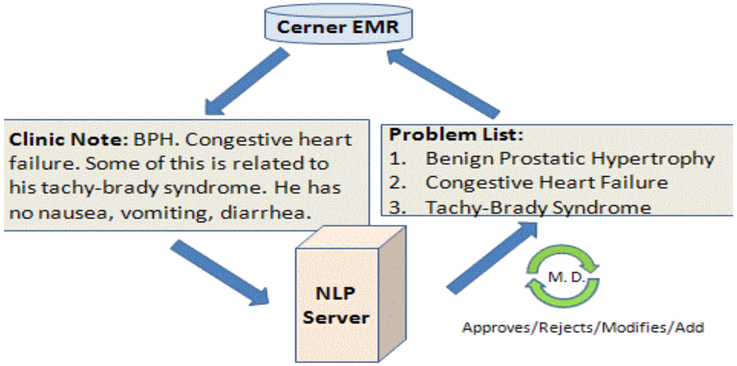
Points of Intervention for NLP Systems



Patient-Centered Clinical Trial Eligibility Screening

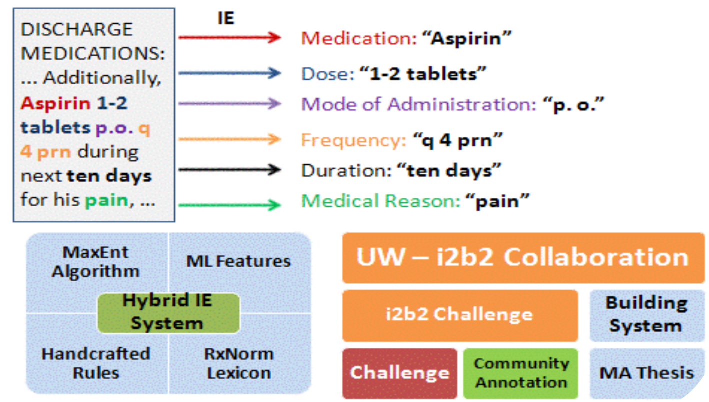
Summary – Questions?

Semi-Automated Medical Problem List



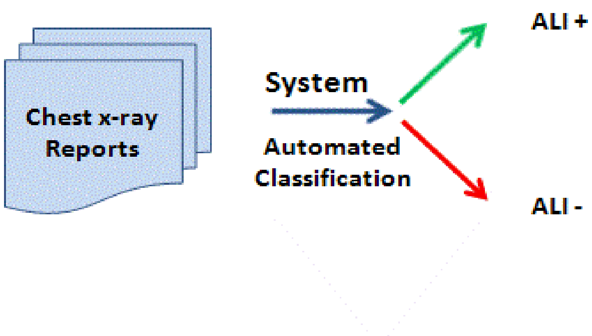
lti I., et al. "Building an automated problem list based on natural language processing:
ns learned in the early phase of development. "AMIA Annu Symp Proc. 2008 Nov 6:687

Automated Extraction of Medication Information

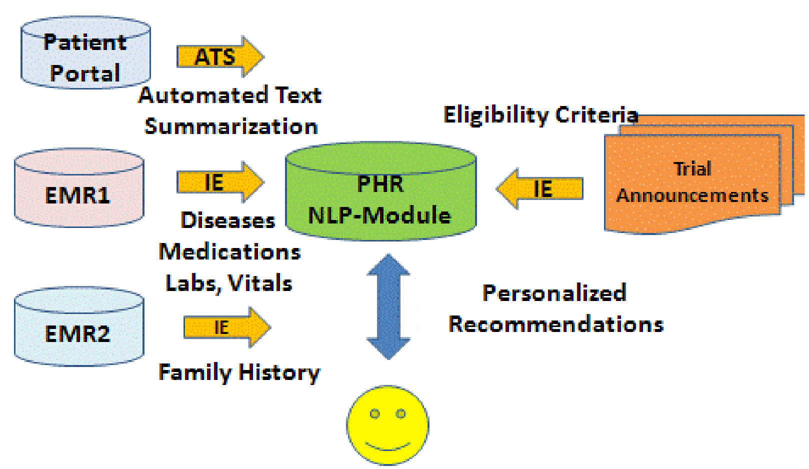


NLP Use Cases for Clinical Informatics and Translational Informatics

Task for Automated ALI Classification



Points of Intervention for NLP Systems



Patient-Centered Clinical Trial Eligibility Screening

Summary – Questions? (Text Version)

<u>Past Projects:</u>

1. Semi-Automated Medical Problem List: Clinical-NLP, IE, NER - 1 Slide
2. Extraction of Medication Information: Clinical-NLP, IE, NER - 1 Slide
3. **Classification of Radiology Reports for Acute Lung Injury**: Clinical Document Classification

<u>Future Project:</u>

4. ***Automated Clinical Trial Eligibility Screening**: Clinical NLP, Biomedical-NLP, IE, NER, Document Classification

***Grant funded**