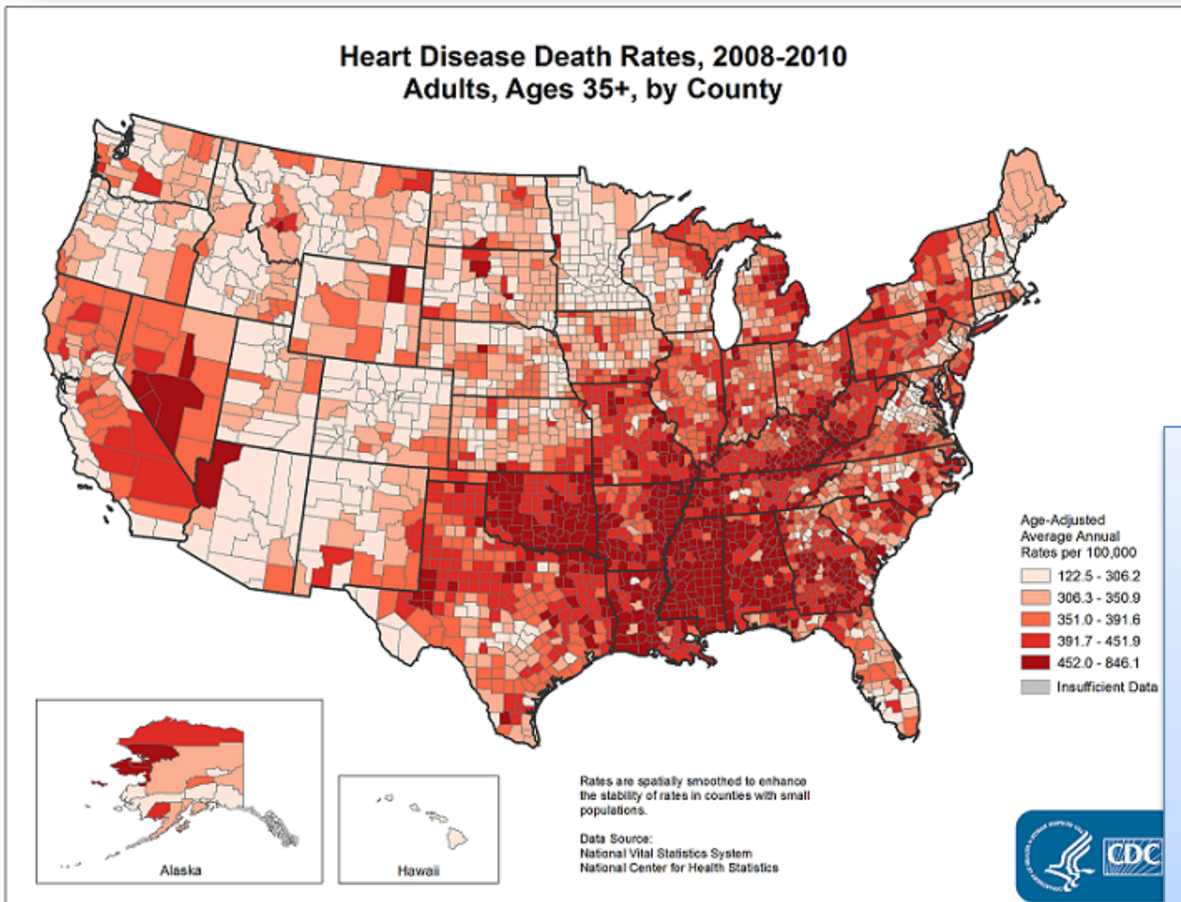
TMU UNSW

TMUNSW System for Risk Factor Recognition and Progression Tracking

Nai-Wen Chang[1,2], Hong-Jie Dai, PhD[3*], Chih-Wei Chen, MD[3], Jitendra Jonnagaddala[4], Chou-Yang Chien[5], Manish Kumar[4], Richard Tzong-Han Tsai, PhD[5], Wen-Lian Hsu, PhD[1]

[1]Institution of Information Science, Academia Sinica, Taiwan; [2]Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taiwan; [3]Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, Taiwan; [4]Translational Cancer Research Network, University of New South Wales, Australia; [5]Computer Science and Information Engineering, National Central University, Taiwan

Heart disease is the number one cause of death for both men and women in the United States.



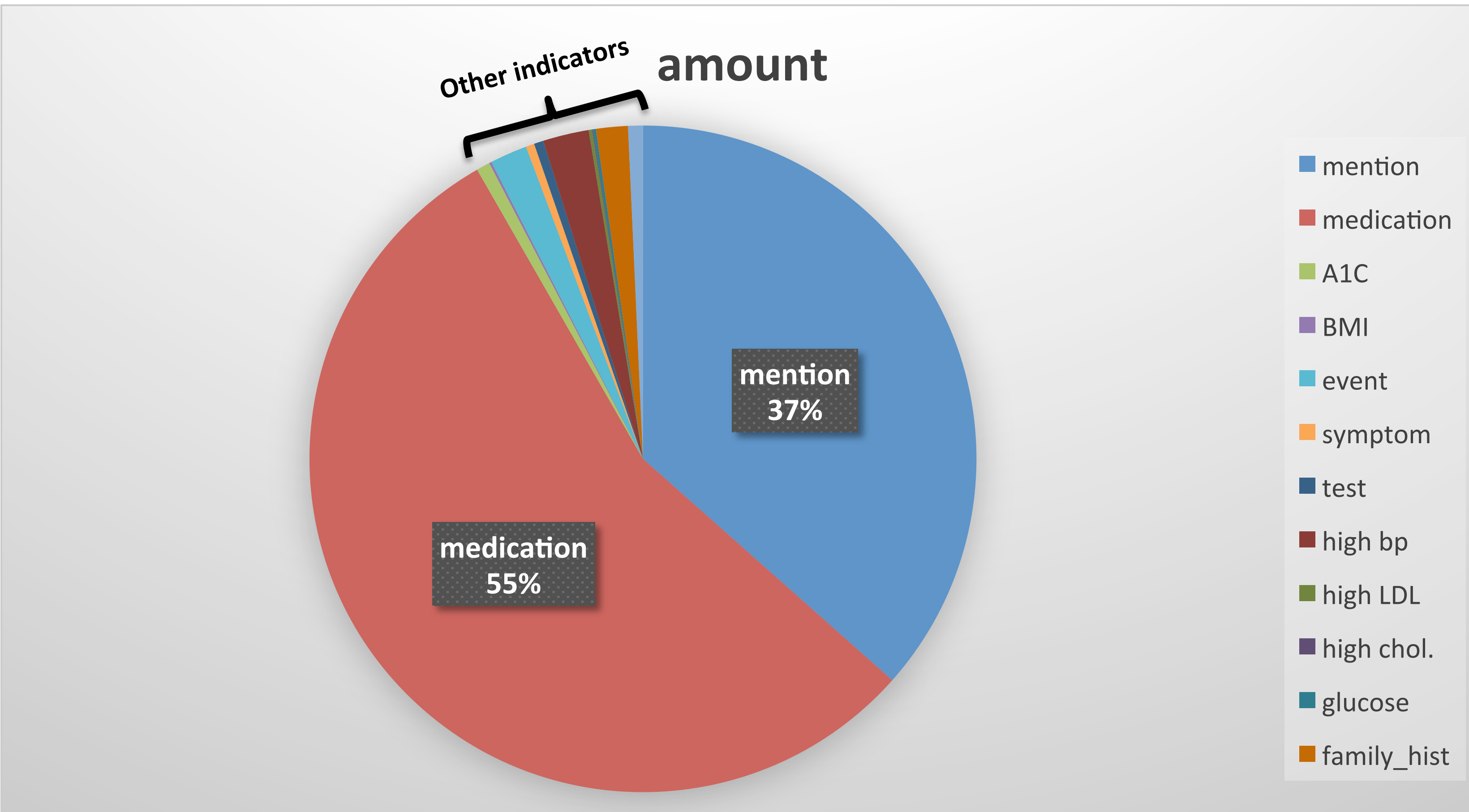
Race of Ethnic Group	% of Deaths
African Americans	24.5
American Indians or Alaska Natives	18.0
Asians or Pacific Islanders	23.2
Hispanics	20.8
Whites	25.1
All	25.0

- 600,000 people/year die of heart disease
- More than 50% of the deaths were in men in 2009
- 380,000 people are also Coronary heart disease
- 720,000 Americans have a heart attack
- Coronary heart disease alone costs the United States $108.9 billion each year

http://www.cdc.gov/heartdisease/facts.htm

TMUNSW:
A **context-aware approach** to assign the time attributes for all recognized medical concepts

Distribution of data type



Methods

Pre-processor

- Split sentence
- Tokenization
- Stemming
- Removing stop words

Concept recognizers

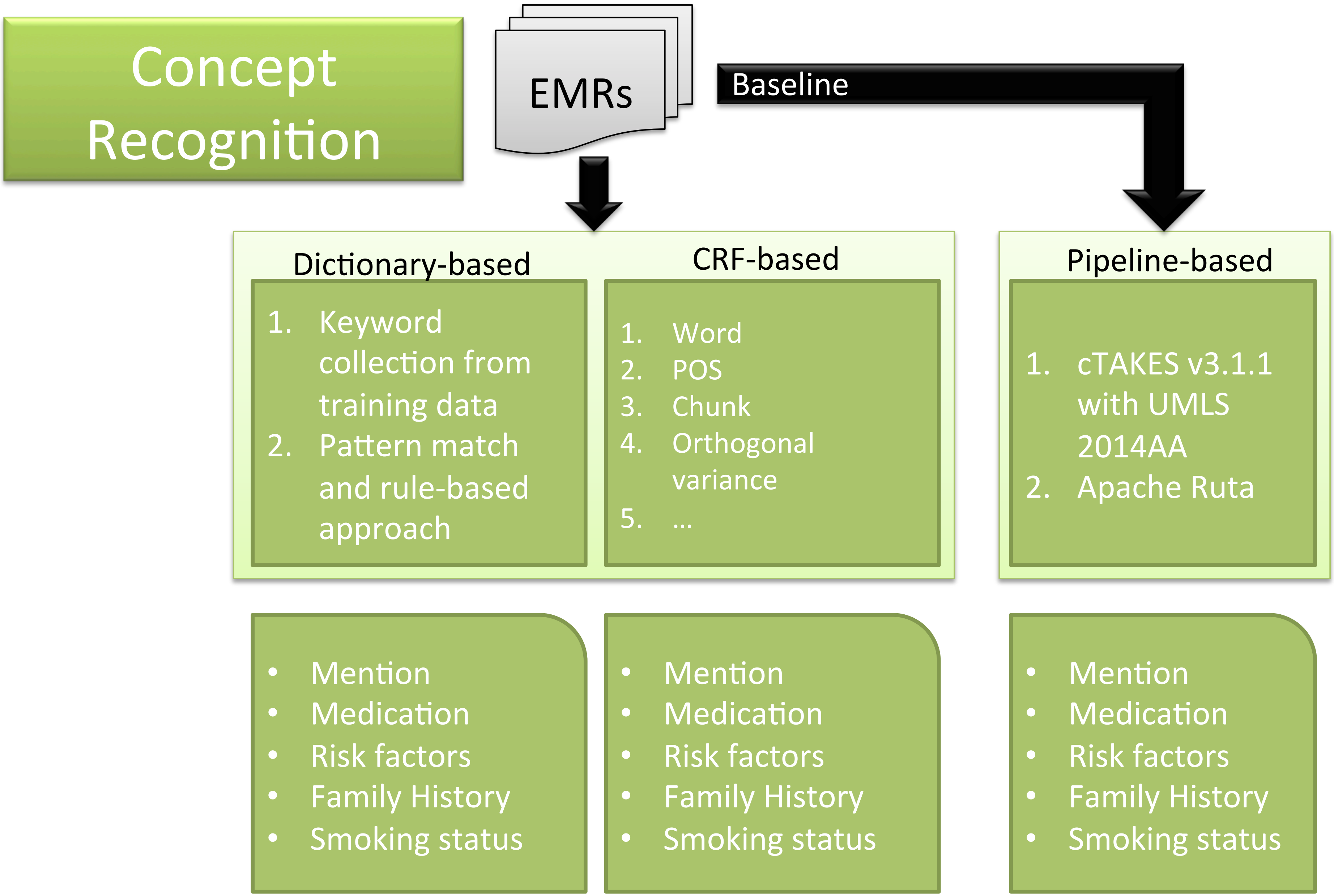
- **Disease mentions**
- **Corresponding risk factors**
- **Medications**
- …etc

Status Classifier/ Time-attribute Assigner

- **Section Recognizer**
- Status classifier
- Time-attribute assigner

Concept recognizers

1. The mention concept recognizer
2. The risk factors recognizer
3. The medication recognizer


Concept Recognition
EMRs
Baseline
Dictionary-based
1. Keyword collection from training data
2. Pattern match and rule-based approach
CRF-based
1. Word
2. POS
3. Chunk
4. Orthogonal variance
5. …
Pipeline-based
1. cTAKES v3.1.1 with UMLS 2014AA
2. Apache Ruta
• Mention
• Medication
• Risk factors
• Family History
• Smoking status
• Mention
• Medication
• Risk factors
• Family History
• Smoking status
• Mention
• Medication
• Risk factors
• Family History
• Smoking status

The mention concept recognizers

- Dictionary-based recognizer
  - All texts tagged as the “mention” concept within the training dataset were collected and normalized by removing stop words
  - 220 terms were collected for all five mention types
- Machine learning-based recognizer
  - The training dataset annotated with the mention concepts were selected as the training set for the machine learning-based recognizer.
  - Conditional random field (CRF) algorithm was used to build a model to recognize mention concepts.

The risk factors recognizer

1. For each risk factor category, a set of keywords was collected.
2. The list was then used as a dictionary by our system to tag the given medical record.
3. The factor was then reserved for the assignment of time attribute in the later stage

Table 1. Summary of the targeted diseases and their corresponding risk factor definitions

Category	Risk Factor	Numeric Value
Diabetes	High **A1C**	>=6.5
Diabetes	High **glucose**	>126
Hyperlipidemia	High **cholesterol**	>=240
Hyperlipidemia	High **LDL**	>=100 mg/dL
Hypertension	High **blood pressure**	>=140/90 mm/hg
Obesity	**BMI**	> 30
Obesity	Waist circumference	Men:>= 40 inches; Women: >= 35 inches

The medication recognizers

- Dictionary-based and CRF-based approach
- All recognized medications were then matched with a **medication name-category mapping file** to determine the corresponding medication categories.
- The medication terms collected from Wikipedia
  - Generic names
  - Classes of all drugs
- The final dictionary file contains 21 categories and a total of 474 names

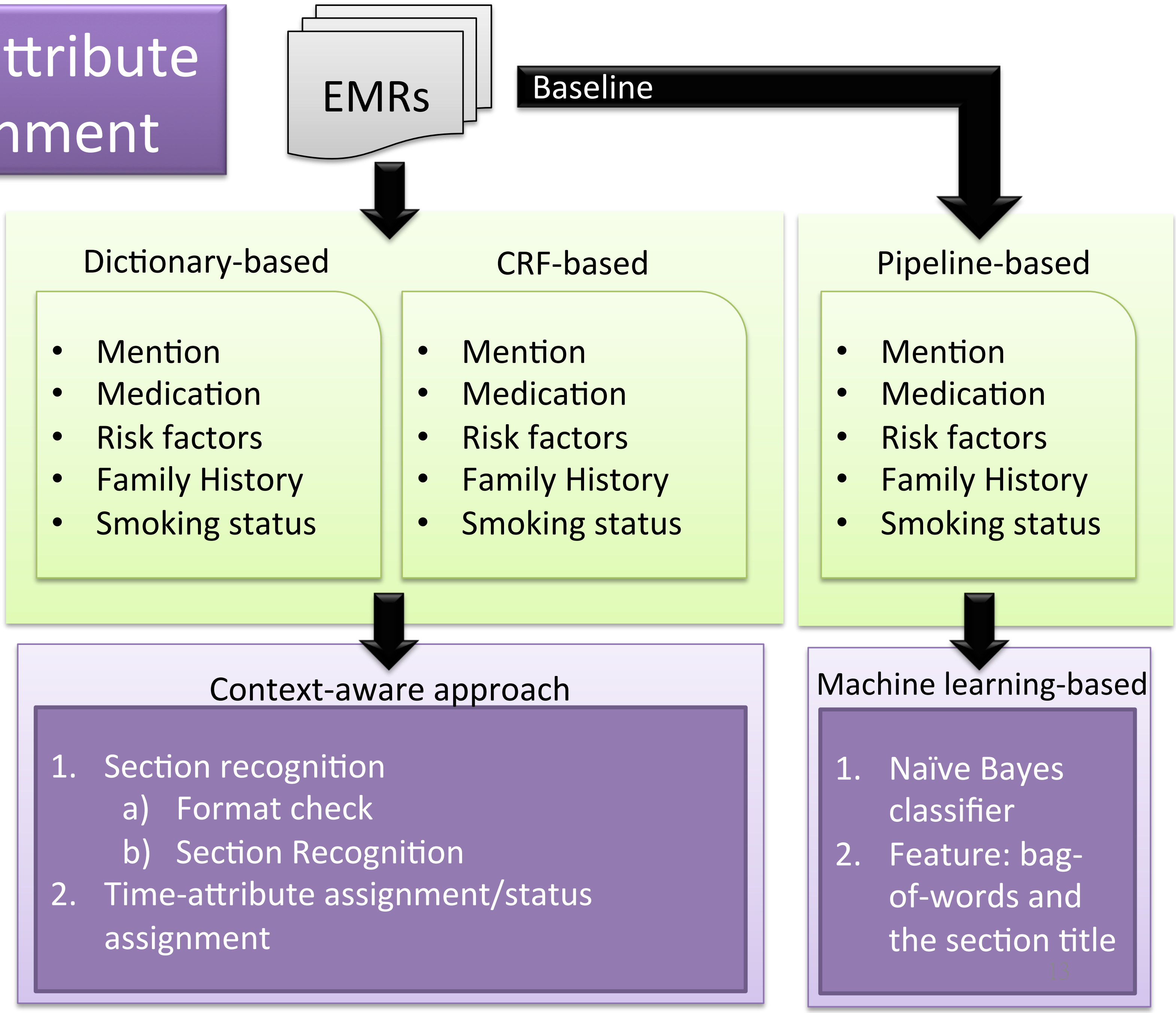
Pipeline-based Medical Concept Recognizers

- A separate clinical document pipeline
  - Apache clinical Text Analysis and Knowledge Extraction System (cTAKES)
    - The Aggregate Plaintext UMLS Processor analysis engine of cTAKES v3.1.1 with UMLS 2014AA as the underline dictionary was employed to recognize medications and the mention concepts.
  - Apache Ruta was used to define patterns that can capture other risk factors, such as blood pressure and HbA1C.

Time-attribute Assigners

1. Context-aware time attribute assignment
2. Machine learning-based time-attribute assignment

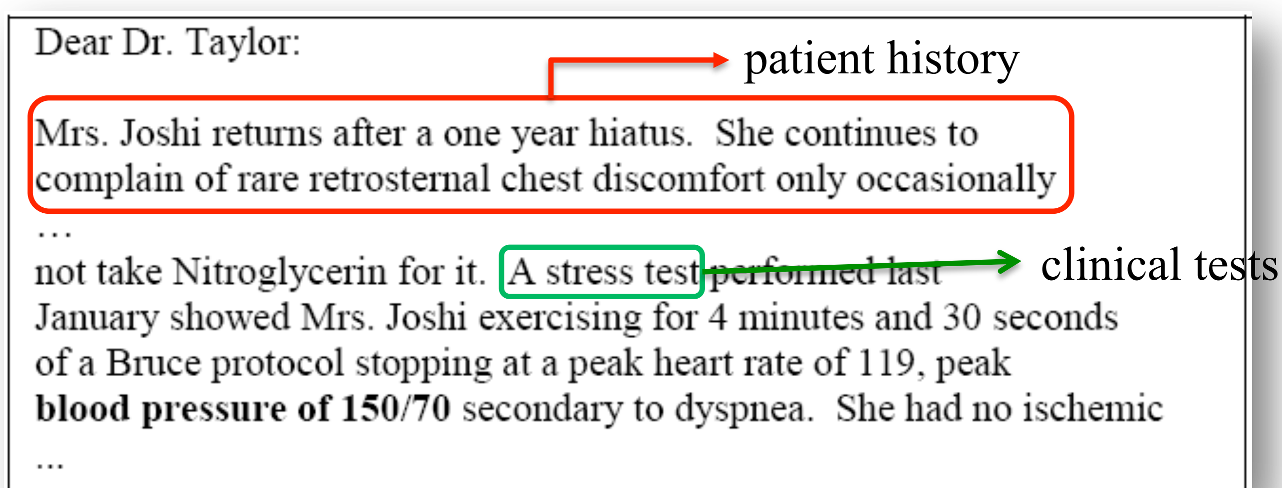
Time-attribute assignment



Electronic medical records (EMRs)

- EMRs facilitate the storage, retrieval, and exchange of the health information of an individual patient.
- Information are stored in the form of **free text** within the EMR.
- Two data format: Email and Section

Email format


Dear Dr. Taylor:

patient history

Mrs. Joshi returns after a one year hiatus. She continues to complain of rare retrosternal chest discomfort only occasionally

…

not take Nitroglycerin for it. A stress test performed last January showed Mrs. Joshi exercising for 4 minutes and 30 seconds of a Bruce protocol stopping at a peak heart rate of 119, peak **blood pressure of 150/70** secondary to dyspnea. She had no ischemic

clinical tests

…

Electronic medical records (EMRs)

Section format

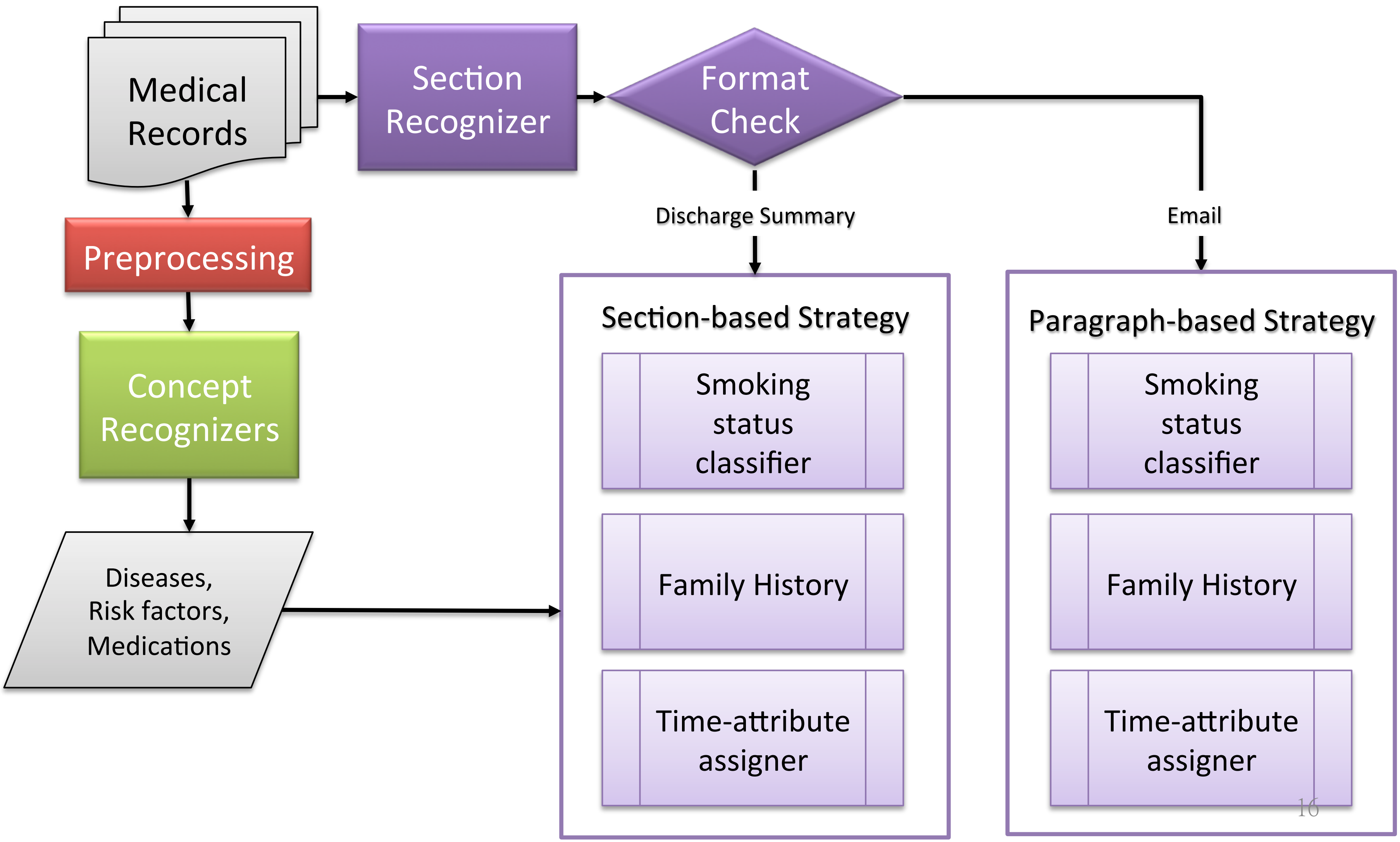
Inaccessible and infeasible for searching, summarization and analysis

Record date: 2137-02-27
CARDIOLOGY
PACIFIC COAST HOSPITAL
Reason for visit:
transfer from Colorow, chest pain in setting of known CAD
Interval History
63-year-old woman with multiple medical problems, notably CAD, s/p RCA and LCx PCI in the context of NSTEMI in March, 2136, with return to PCH in July, 2136 with recurrent chest discomfort. Cardiac catheterization during that visit revealed in-stent stenosis in LCx stent, successfully addressed with bare metal stent placement. Most recent nuc. stress 10/03/36 showed no definite ischemia, mild apical and mild inferolateral thinning not clearly outside normal per report.

…

Past medical history:
1. CAD, s/p NSTEMI in March, 2136 with BM stent placement to LCx and RCA
Coronary angiogram in July, 2136:
2. Hyperlipidemia
3. Hypertension
4. MAC, previous treatment course terminated due to hepatitis in 2128
5. Bronchiectasis, s/p right middle lobectomy in 2128

Context-aware time attribute assignment



Paragraph-based strategy

Email format

Preprocessing

Time-attribute assignment

Concept Recognize

Dear Dr. Taylor:

Mrs. J
comp
…
not ta
Janua
of a B
blood
…

2.4 Hypertension

indicator	description
mention	a diagnosis of Hypertension or a mention of a pre-existing condition
high blood pressure	BP measurement of over 140/90 mm/hg (if either value is high, the patient has hypertension)

Table 4: indicators and descriptions for the Hypertension tag

Dear Dr. Taylor:

Mrs. Joshi returns after a one year hiatus .
She continues to
complain of rare retrosternal ch
…
not take Nitroglycerin for it .
A stress test performed last
January showed Mrs. Joshi e
seconds
of a Bruce protocol stopping at a peak heart rate of 119 , peak
blood pressure of 150 / 70 secondary to dyspnea .
She had no ischemic
…

Sentence re-combination

A stress test performed **last January** showed Mrs. Joshi …, peak **blood pressure of 150/70** secondary to dyspnea.

Before DCT

High bp

Section-based strategy

Section format

ate: 2137-02-27
CARDIOLOGY

Record date: 2137-02-27
CARDIOLOGY
PACIFIC COAST HOSPITAL
Reason for visit:
transfer from Colorow, chest pain in setting of known CAD
Interval History:
63-year-old woman with multiple medical problems , notably CAD , s / p RCA and LCx PCI in the context of NSTEMI in March , 2136 , with return to PCH in July , 2136 with recurrent chest discomfort .
Cardiac catheterization during that visit revealed in-stent stenosis in LCx stent , successfully addressed with bare metal stent placement .
Most recent nuc .
stress 10 / 03 / 36 showed no definite ischemia , mild apical and mild inferolateral thinning not clearly outside normal per report .

CAD and before DCT

Machine learning-based time-attribute assignment

- A machine learning model based on the naïve Bayes classifier was build to assign the time-attribute of mentions and integrated it into the developed UIMA pipeline.
- The employed features included bag-of-words and the section title information.
- The three time attributes, including “during DCT”, “before DCT”, and “after DCT” were used as class labels.

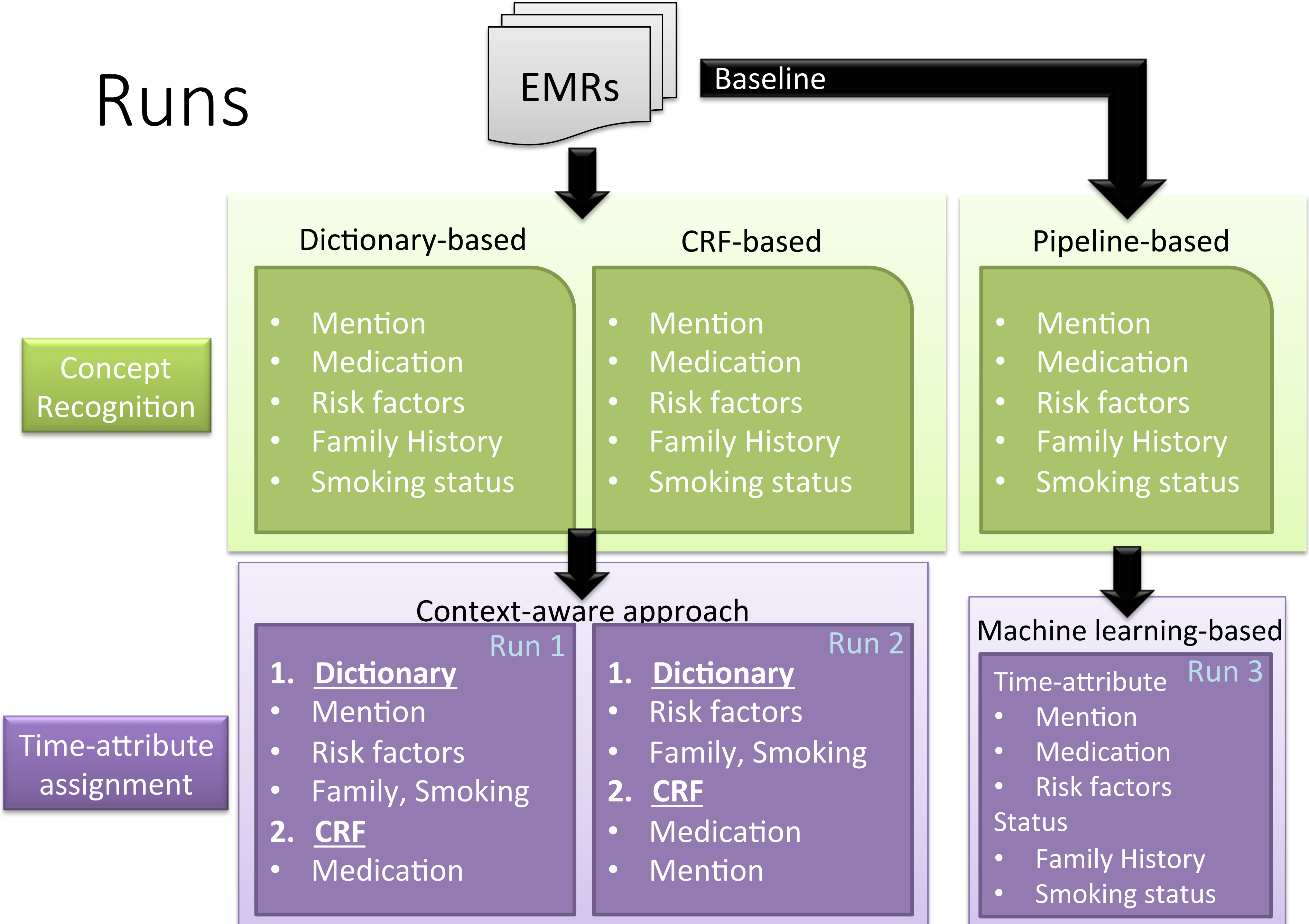
Family History Status Classifier

- Dictionary-based tagger
- The family history status (“present" or "not present“)
- **First**-degree relative (parents, siblings, or children)
- diagnosed prematurely
  - Age > 55 for male relatives with CAD
  - Age > 65 for female relatives with CAD
- “Present”
  - The sentence contains a male/female first-degree relative name, along with specific age-related information
  - The sentence contains CAD-related terms and even numbers of negation terms.

Smoking Status Classifier

- Dictionary-based tagger
  - **Smoking-related keywords**(“smoking”, “cigarette”… etc)
- The text containing the listed terms was regarded as the context of the smoking status, and several weighted rules developed for different smoking statuses were applied on the context to decide the smoking status of the patient.
- If the context did not provide sufficient information to determine the smoking status
  - **Context-aware approach**

Runs


EMRs
Baseline
Concept Recognition
Dictionary-based
• Mention
• Medication
• Risk factors
• Family History
• Smoking status
CRF-based
• Mention
• Medication
• Risk factors
• Family History
• Smoking status
Pipeline-based
• Mention
• Medication
• Risk factors
• Family History
• Smoking status
Time-attribute assignment
Context-aware approach
Run 1
1. Dictionary
• Mention
• Risk factors
• Family, Smoking
2. CRF
• Medication
Run 2
1. Dictionary
• Risk factors
• Family, Smoking
2. CRF
• Medication
• Mention
Machine learning-based
Run 3
Time-attribute
• Mention
• Medication
• Risk factors
Status
• Family History
• Smoking status

Result

1. dictionary-based mention recognizer
2. machine learning-based medication recognizer
3. machine learning-based mention recognizer
4. proposed context-aware approach for the assignment of time attribute.

- For the **machine learning-based** system, a **10-fold cross validation** on the same dataset was applied to select efficient features for mention concept and medication recognition.
- Run 1 = 1+2+4
- Run 2 = 2+3+4
- Run 3 = pipeline-based method that is entirely based on the machine learning approach, including the recognition of mention and medication and the assignment of time attribute.

Table. The performance of each submitted run of the TMUNSW system and the aggregated results of all runs.

	Run1	Run2	Run3	Mean	Median
Micro Precision	0.8594	0.8384	0.621	0.808	0.852
Micro Recall	0.9387	0.9404	0.6562	0.835	0.908
Micro F1	0.8973	0.8865	0.6381	0.815	0.872

Conclusion

- TMUNSW is a system that can recognize the concepts such as medication, risk factors and diseases and track the progression with time-attribute assigner.
- We have released our tool: tmuClinical.NET for researchers and users as the link: https://sites.google.com/site/hongjiedai/projects/tmuclinicalnet
- The context-aware assignment approach outperforms machine learning-based and rule-based w/o awareness of context.

tmuClinical.NET

https://sites.google.com/site/hongjiedai/projects/tmuclinicalnet

Projects >

tmuClinical.NET

tmuClinical.NET is a set of C# library developed for processing discharge summaries.

About
Downloads
Developers

tmuClinical.NET was built using Microsoft .NET Framework 4. Its components were developed based on the data set of the i2b2 2014 shared task track 2.

tmuClinical.NET employs a number of rule-based and machine learning methods. It also integrates several state-of-the-art natural language processing components available on the NuGet Gallery.

tmuClinical.NET components include:

- Section heading recognition
- Risk factor recognition and time attribute assignment
- Smoking status classification
- Record format detection
- Tokenization, part-of-speech tagging and sentence boundary detection through MedPost

Any Questions?

Contact: nwchang@iis.sinica.edu.tw

Nai-Wen Chang

Thank you