1 Glossary

- **Annotation**: Annotation systems are the markup styles used by NLP systems (called machine annotation) and by humans (called human annotation) to highlight target NLP mentions or other features in a text, like the start/stop of a sentence, an NER term, or a part-of-speech. When an NLP system annotates text this is called labeling ("to label a mention").

- **Assert**: An instance of NLP system output.

- **Assertion**: noun Assert.

- **CLAMP**: Clinical Language Annotation Modelling and Processing tool kit.

- **Context**: In health NLP, context is means whether a term mentioned in a clinical text is a term of interest to an NLP system. These are the most important context issues in health NLP:
  Negation: if the term is excluded, e.g., "The patient does not have cancer."
  Experiencer: if the term refers to the current patient or to someone else, e.g., "Her father had cancer."
  Historical: if the term applies to the current note or to the past. "The patient had cancer three years ago but now is cancer free."The first instance of "cancer"is historical, the second instance of "cancer"is a negation example.
  Probable: if the term is uncertain or not confirmed, e.g., "The patient may have cancer but we need more testing to be sure."

- **Corpus (plural: Corpora)**: A collection of texts processed by an NLP system. Examples of a corpus in health NLP:
  All of the clinical notes belonging to a patient (small corpus).
  All of the notes belonging to an entire hospital (big corpus).
  All articles ever written on natural language processing (very big corpus).

- **FN**: False negative.

- **FP**: False positive.

- **Framework**: An NLP framework is specific approach to NLP processing. Typically, a framework is defined by its underlying tool set and its annotation system. UIMA, GATE, and CLAMP are examples of NLP framework.

- **GATE**: GATE, Generalized Architecture for Text Engineering, is a Java-based open-source NLP framework and tool set created at the University of Sheffield. It is one of the most mature frameworks available. Its use is not limited to clinical applications, but it has no special tools for clinical NLP

- **Healthcare system**: All hospitals and clinics. The Vietnamese healthcare system is all hospitals and clinics in the country.

- **Machine Learning**: Machine learning, abbreviated ML, is programming computers so they can learn from data. Simple to define, hard to do!

  Machine learning is used in two ways in health NLP:

  Some NLP systems use ML to help find mentions in text. They might learn, for example, that every time the pattern <number>"C"occurs in a clinical note, the <number> means a temperature value. "John had a temperature of 40C and was sweating."

  After finding mentions in text, ML is often used to make higher level assertions at the note, encounter, or patient level. For example, if "fever"is asserted multiple times in a patient's collection of notes, then the system might assert that the patient has "chronic fever."That is an example of a patient-level assertion.

- **Mention**: An occurrence of term of interest in one or more texts being analyzed by an NLP system. In health NLP, a "mention-= an NLP "target".

- **NER**: Stands for "named entity recognition."NER is an NLP task that finds mentions of specific terms of interest to a specific NLP project. Clinical NER includes medical terms not usually found in everyday texts. The most common types of medical terms are: medication names (like "penicillin"or "paracetamol"), allergies (like "reaction to penicillin"or "pollen allergy"), a diagnosis (like "heart attack"or "stomach cancer"), a symptom (like "fever"or "dizzy"), and medical tests (like "serum glucose"or "x-ray").

  Clinical NER is complicated by the use of short forms like acronyms (like "MI"for "myocardial infarction"or "ECG"for "electrocardiogram") and abbreviations (like "glu"for glucose or "pt"for patient).

- **Note**: A description of a patient encounter written by a nurse, doctor, or other member of the healthcare system.

- **Parsing**: Parsing is an NLP task that extracts part-of-speech, or chunks of speech, from texts. The VnCoreNLP system is an example of a Vietnamese parser. The Stanford Parser is an example of an English parser (one of the best).

- **Patient encounter**: An episode of care.

- **Performance metrics**: These are standard ways to determine how well an NLP system is working. They are all based on comparing the NLP output of your system to a reference (or gold) standard. The reference standard is frequently built from human annotations.

  For example, assume your NLP system is processing 1,000 texts. And assume that in those texts there are 100 clinical-term mentions of interest to your project (like 100 diagnoses and your system is trying to find all diagnosis mentions). To run a performance metric, a human has to annotate these mentions. These 100 mentions are called "true positives (TP)". Also assume that there are 200 other clinical terms that are not diagnoses. These are called "true negatives (TN)". When your system labels a TN as a positive, this is called a "false positive (FP)". When your system labels a TP as a negative (in this example, a non-diagnosis term is labeled as diagnosis), this is called a "false negative (FN)". Once you understand TP, TN, FP, and FN, you have everything you need to compute most important performance metrics.

  The most popular metrics are:

  - Accuracy: fraction of all the labels assigned by your system that are correct.

    $$\frac{TP + TN}{TP + TN + FP + FN}$$

  - Accuracy can be useful, but if TP are rare, then simply labeling every mention as a negative mention gives a high accuracy.

  - Precision: the fraction of all your system's positive labels that are really positive.

    $$\frac{TP}{TP + FP}$$

  - Precision goes up as false positives go down. This is also called "positive predictive value."

  - Recall: the fraction of all the positive mentions you should have found that were labeled as positive.

    $$\frac{TP}{TP + FN}$$

    Recall goes up when the false negatives go down. This is also called "sensitivity."

  - F-score (or F-measure): a combined measure that balances precision and recall equally.

    $$\frac{2 \times Precision \times Recall}{Precision + Recall}$$

  Why is F-score a useful measure? The closer recall and precision are to each other, the more like a simple average this measure becomes. But the more recall and precision differ, the smaller the F-score.

This penalizes a system that is good at precision or recall but not good at both. Technically, this is the "harmonic"mean of precision and recall. Note that some NLP applications might want to weigh recall higher than precision (where it is important not to miss positives) or weigh precision higher than recall (where it is expensive to label a positive in error).

There are many other metrics, and there are ways to adjust F-score to weigh recall and precision differently

- **Pipeline**: A pipeline is an approach to NLP that uses a series of processing modules to analyze texts. Each module performs one task, like sentence segmentation or POS parsing. Typically, a module takes input from the module before it and delivers output to the module after it. This is the most common framework used in clinical NLP today. The modular design allows for quick updates (i.e., to change a parser you need only change the parsing module). Some pipeline frameworks, like UIMA-AS, support parallel stages. Normally this is used to speed up a stage that is a performance bottleneck. CLAMP is the pipeline we use in this class.

- **POS**: An acronym that stands for Part Of Speech. A part-of-speech is a grammatical concept of human speech. Human languages have a general syntax. That syntax is not as rigorous as a computer language. Parts-of-speech are connected by an underlying grammar unique to each language. Human languages share parts-of-speech like verbs (action words) and nouns (things). Typically, POS are extracted from text by a parsing the text.

- **Segmentation**: To break a text into individual sentences (sentence segmentation) or sentences into words or phrases (phrase segmentation).

- **TN**: True Negative.

- **TP**: True Positive.

- **UIMA**: Unstructured Information Management Architecture, a framework from IBM that can be used for NLP. UIMA is not limited to NLP, however. A key aspect of UIMA is its annotation type system. Each stage of a UMA pipeline performs one (typically) kind of annotation, like POS tagging, storing the annotations separately from the text. Input texts are unchanged as they are passed along the pipeline stages. IBM used a version of UIMA, UIMA-Asynchronous Scaleout, in its Watson system. Watson is used in health NLP, but many of IBM's claims for this system are not well supported.

- **VnCoreNLP**: This is an NLP pipeline designed for the Vietnamese language. Inspired by Stanford's

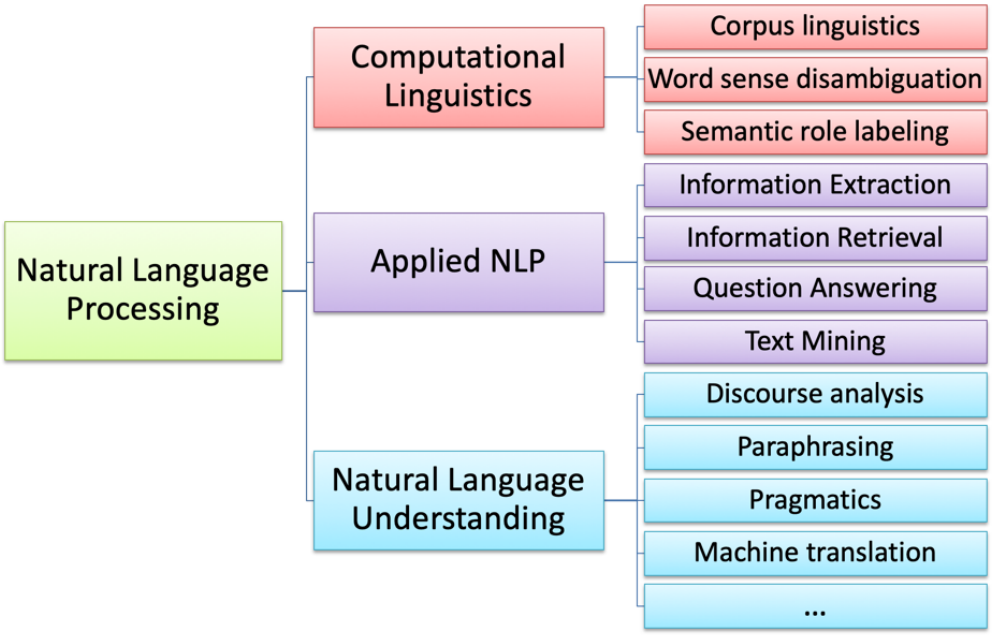
CoreNLP tools, VnCoreNLP offers word and sentence segmentation, POS tagging, and NER – all designed specifically for Vietnamese texts.

2 Week 1 - INTRODUCTION

- **BIG MARKET**
  - VN Health Spending = 32.000.000.000 $USD
  - USA Health Spending = 17.3 % GDP (several trillion 10^{12})
  - 57.4% Vietnamese use cellphone. 14.1% used health App, on 14.1% who did, 66.4% reported apps useful.
- **Some Vietnamese mHealth Apps**
  - A majority of the initiatives targeted vulnerable and hardto-reach populations.
  - Aimed to prevent the occurrence of disease.
  - Used text messaging (short message service, SMS) as part of their intervention.
- **HEALTH NLP HARD/DIFFICULT/CHALLENGING?**
  - Hard.
  - NLP is popular in USA.
  - Few programmers/data scientists specialize in health NLP (less competition for you)
  - Health apps == excellent long-term growth.
- **Why Health NLP is Hard?**
  - Context is critical.
  - Sentence segmentation is critical.
  - Many short forms (not real words in medical dictionary).
- **Trends in Health NLP and Health Data Sciences.**
  - Health NLP started like all NLP. Computional linguistic.
    - Formal language theory: important and essential for all CS majors.
    - Grammars and Parsing: important for formal language theory.

- **FORMAL LANGUAGE THEORY != CLINICAL TEXT**
  - Not too useful with clinical texts.
  - Clinical text has no proper syntax (50% not real sentences).
  - Clinical text has many short forms.
  - BUT linguistic approach still useful in health NLP sometimes for feature extraction.
- **Health NLP BIG PICTURE**

  We only focus on the second part



- **Health data science top skill**
  - Programing: Python and R.
  - Databases: SQL and Hadoop
  - Machine Learning
  - Statistic
  - Data visualization

3 Week 2, NLP REVIEW

- **Sample apps**

- The most successful apps work on data from hospitals or groups of hospital, like IBM Watson, open-source EMR (electronic medical record)
- Apps are built locally by a hospital or clinic ... like my team:
  * Find patients where diagnosis was missed.
  * Extract mentions of embolism
  * Find allergies to drugs (why?); find medications in notes
  * Extract drug reactions
  * Retrieve all notes for all patients that match a set of criteria (e.g., have lung cancer and do not smoke)
  * Predict if a patient will return to hospital within a month (import for rural patients; NLP drives ML)
- In USA, over the past 13 years, the number of hospitals using paper-only patient health records has declined from 69 percent to 1 percent1 à huge increase in clinical note data; This will happen in Vietnam, too
- NLP supports everything in indexing

- **SAMPLE MOBILE/CELL PHONE APPS**
  - Show Kardia output
  - MyFitnessPal: overweight BIG prob in US; what are BIG probs here?
  - Step counters (uses accelerometers in phone)
  - Simple diagnosis ("what is wrong with me?")
  - Should I go to doctor? Take my child to doctor? (e.g., Ada app)

- **ML in Health NLP**
  - Some NLP systems use ML to help find mentions in text.
  - ML is often used to make assertions at the note, encounter or patient level.

- **The main tasks of health NLP**
  - The main goal of any Health NLP system is to "assert"
  - Assert == find clinical terms and classify

- BUT one BIG difference with non-health NLP: multi-level classification
- Level
  * Root is Patient
  * A Patient has encounters in hospital or clinic
  * An encounter has one or more notes.
  * Example note:
    · Patient: Hurdle, John F
    · Date: 26/02/2019
    · Findings: **The patient** (Experiencer) is a normal weight Caucasian male who **presents** (Historical) with a distended stomach. He describes "having eaten too much good Vietnamese food" **No fever** (Negation) .
  * Step
    · Assert Mentions
    · Assert context each mention then combine a note's mentions ... assert about **Note**.
    · Combine note assertions ... assert about **Encounter**

- **Processing modules for Health NLP task**
  - Commercial tools: quick start/expensive/good? <demo AWS>
  - Dictionaries <show sample in VN>

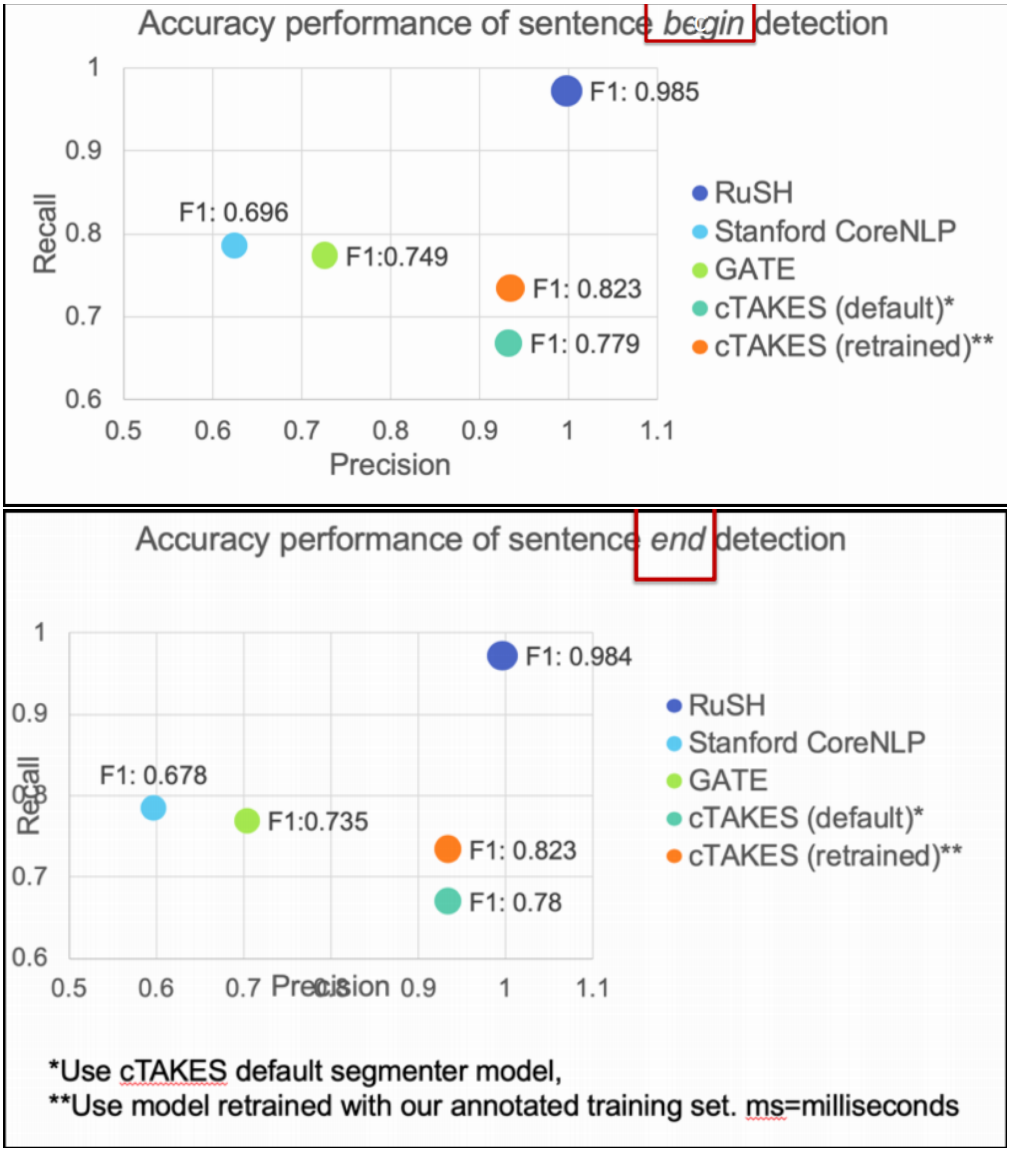
- **Word segmentation in Vietnamese**
  - Clinical dictionary helps word segmentation

Table 2. Word segmentation result

	A	B	C	D
Recall	91.70%	**95.30%**	89.40%	**95.10%**
Precision	88.20%	**94.90%**	85.50%	**93.70%**

A: vnTokenizer (without additional dictionary)
B: vnTokenzier (with additional dictionary)
C: DongDu (without retrained on new data)
D: DongDu (retrained on new data)

- **What is feature extraction**
  - text as data
  - feature vectors
- **How classic NLP useful in Health NLP today**
  - feature extraction
  - Part of speech often useful with other NLP features
- **Sentence segmentation clinical text**
  - Very hard to do with ML: needs Huge corpus

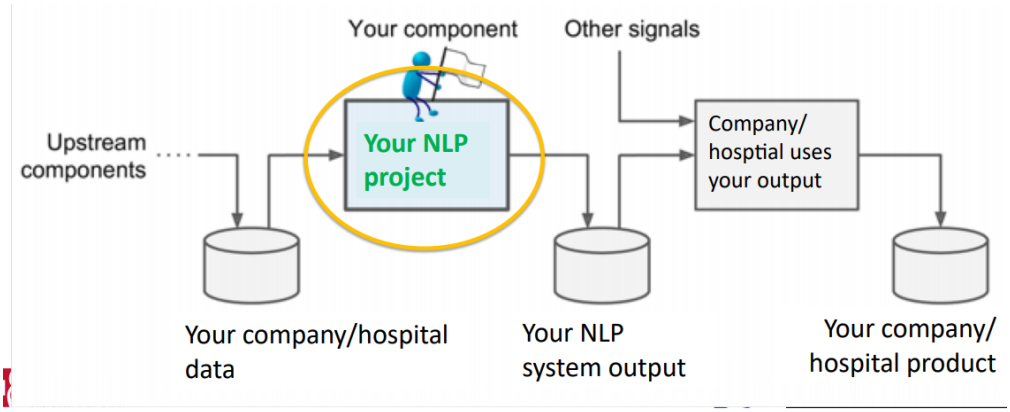


- **Chinese study shown earlier**

- Used 1.360.000 notes
- Chinese had it easy: only around 4000 tokens
- In my research we studied half as many English notes > 100,000 tokens

4 Week3, Health NLP system design

- **OVERVIEW OF NLP OUTPUT SYSTEM DESIGN**
  - Main goal: TEXT INPUT –> NLP OUTPUT.
  - NLP systems are part of a larger system.



  - Using Rules:

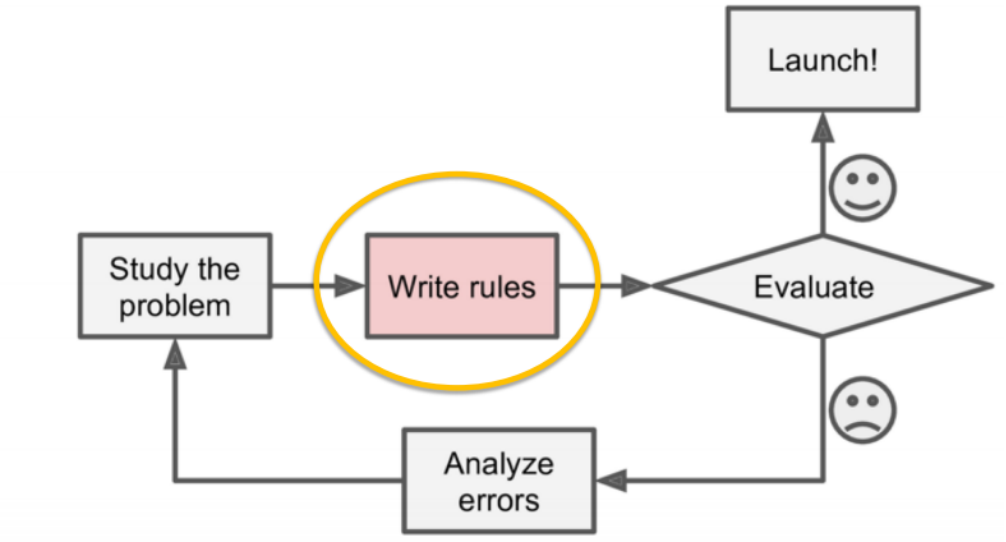


Figure 1-1. The traditional approach

  - Using ML:

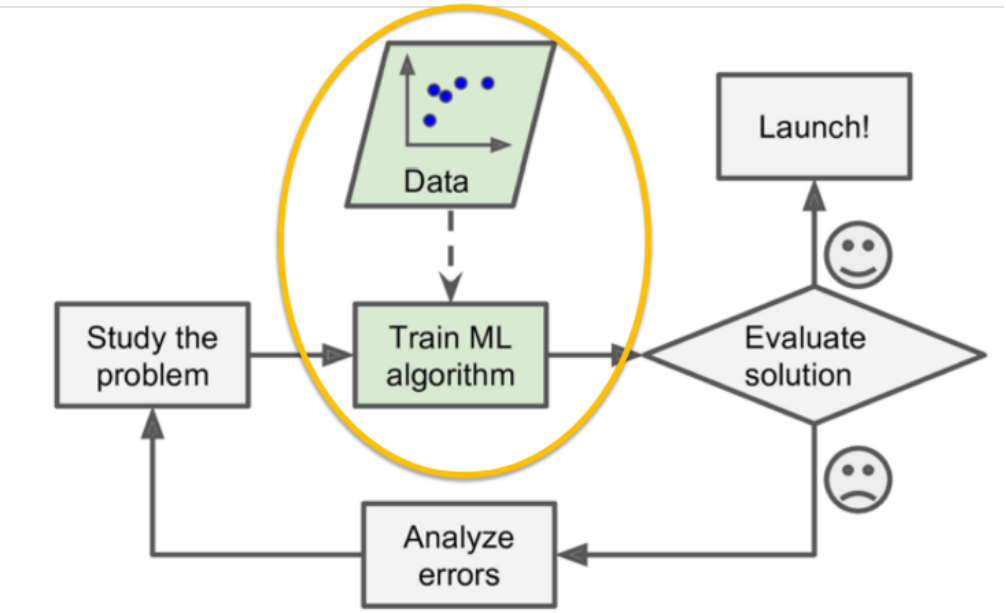


Figure 1-2. Machine Learning approach

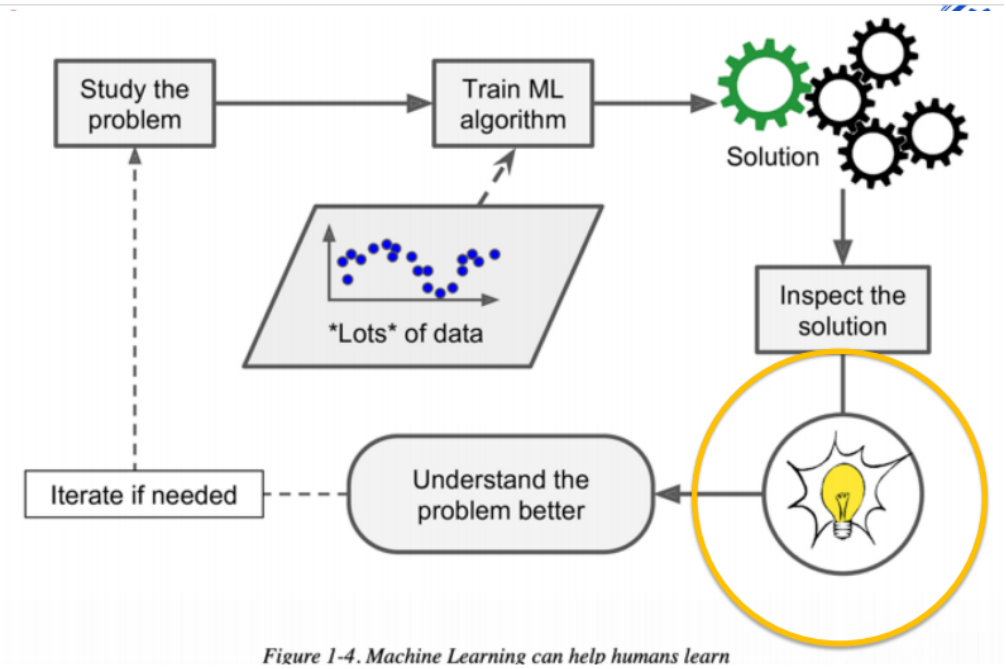


Figure 1-4. Machine Learning can help humans learn

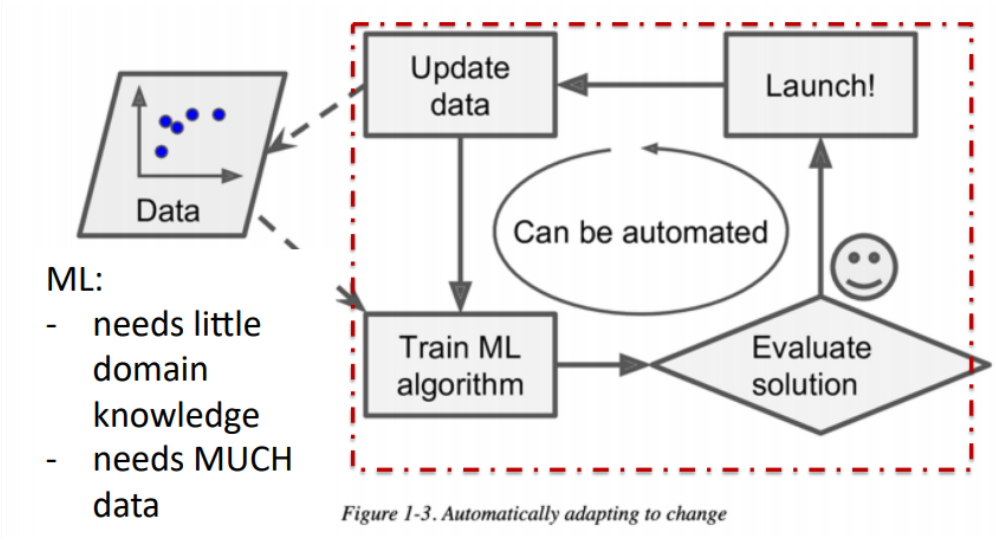
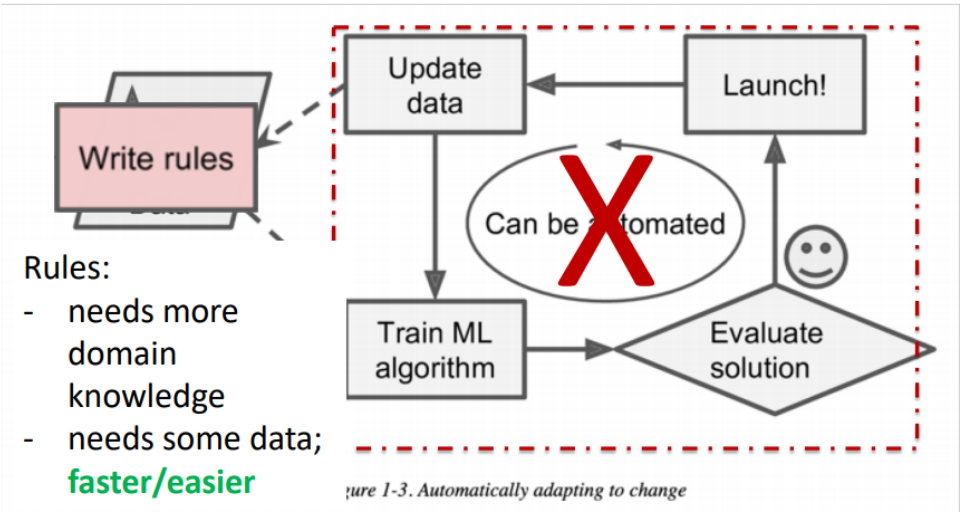


Figure 1-3. Automatically adapting to change

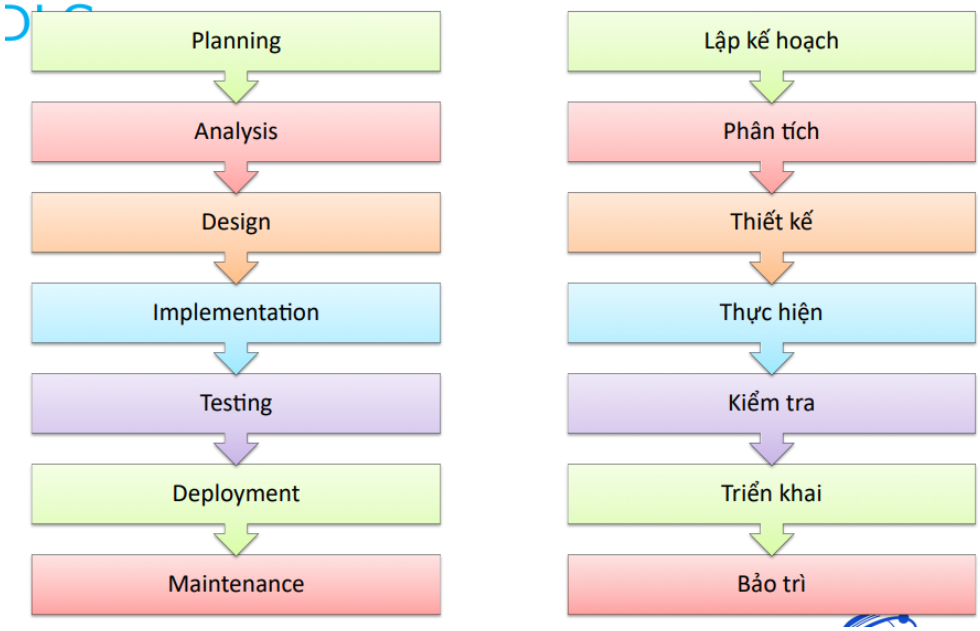


ure 1-3. Automatically adapting to change

– Example of NLP output system. Results come from statistics of words, no rules or part of speech.

- **OVERVIEW OF SYSTEM DEVELOPMENT LIFE CYCLE (SDLC)**

  – Seven steps that can be used to build any real-world system

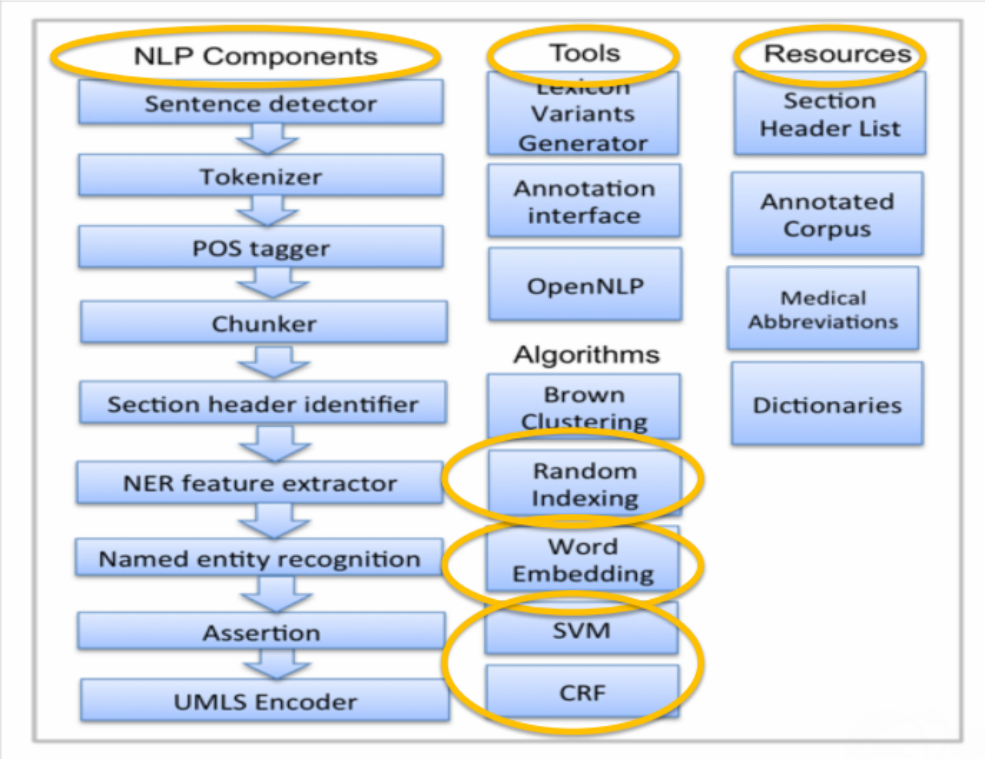


  - Planning:
    - System planning: what is the goal of the extraction system.
    - Define target mentions: formal definition of NLP targets (mentions) into measurable factors.
  - Analysis:
    - Esimate how feasible to extract mentions.
    - Are target mentions very ambiguous?
    - How hard is sentence segmentation
    - The health NLP goal maybe to hard
  - Design:
    - Build NLP system to meet analysis plan.
  - Testing

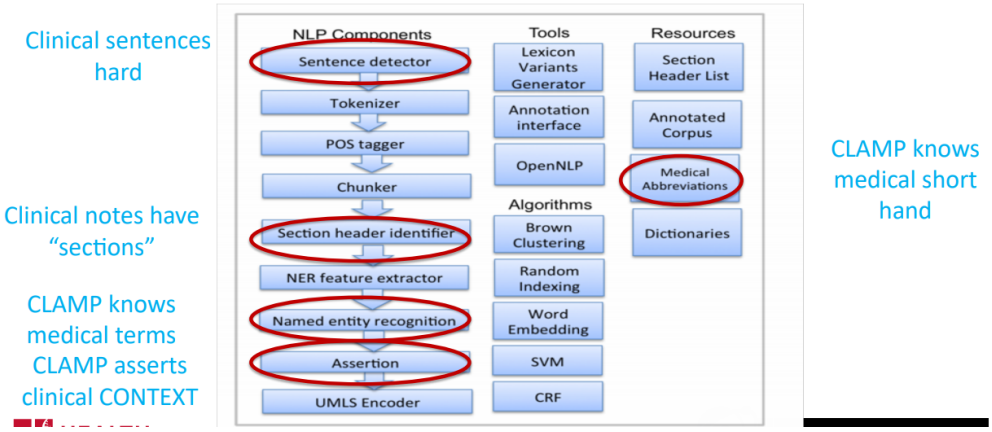
- **NLP reference standard**
  - Annotation
  - Symantics
  - Testing: Human annotations vs NLP annotations

5 Week 4, CLAMP

- CLAMP Pipeline modules do all basic NLP.



- CLAMP pipeline modules do hard NLP



- CLAMP using GUI
- CLAMP using Rules and ML

6 Week 5, CLAMP 2 and Projects

- What are annotations:
  - Annotataion is critical for testing and ML
  - When you annotate, you see patterns that == rule logic or rule-based system

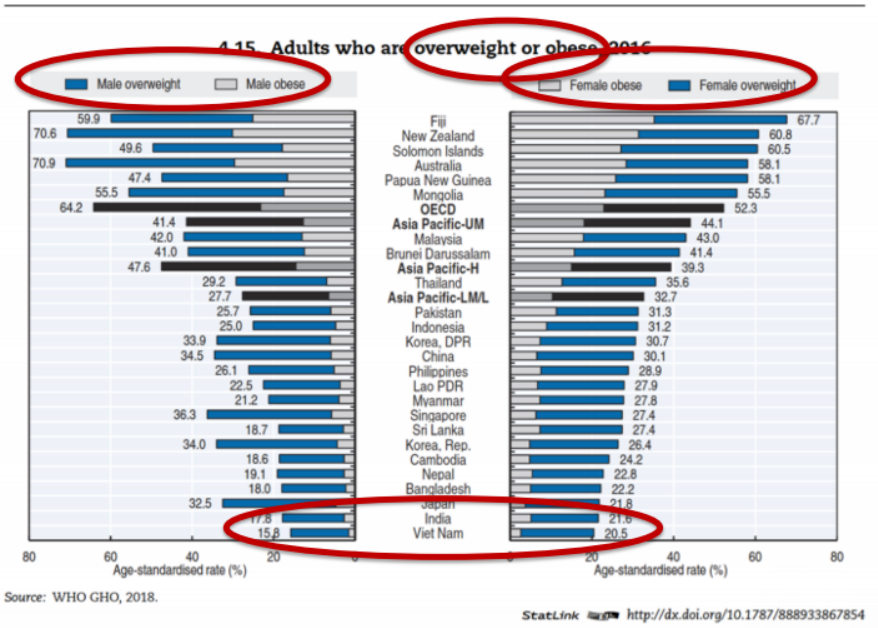
7 Week 6, Project and CLAMP examples

8 Week 7, DEEP DIVE INTO CLAMP NLP

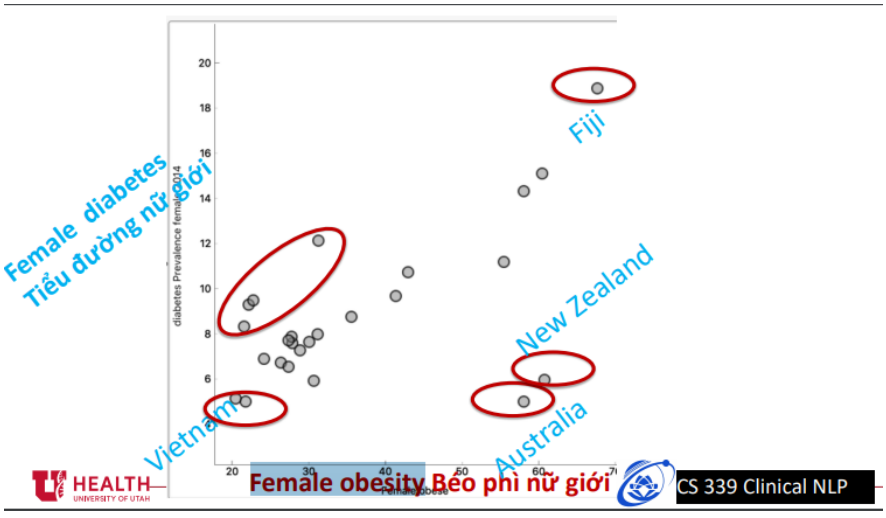
- SENTENCES ARE ALWAYS THE FIRST PIPELINE COMPONENT

9 Week 10, NLP AND DATA MINING TOOLS AND RESOURCES

- ASIAN OBESITY AND DIABETES DATA



- Female obesity



- UMLS: Unified Medical Language System

  - Created by US National Institutes of Health (NIH)
  - NIH mapped 100+ different medical dictionaries to a common code, the code is called CUI (concept unique identifier).

10 Week 12, INTRODUCTION TO MACHINE LEARNING USING ORANGE3

- Why use O3 at all? You can program your own ML...
  - Every ML project needs a baseline.
  - Lower bound on performance
  - Your ML system must do better than baseline.
  - Quick start. No programming overhead.
  - Easy to spot trends or data problems early.
  - Good for SDLC Analysis step: is project possible?
  - Whole team can browse ML models at same time (improves sharing, synergy, and communication)
  - Easy for you to focus on data. Need more data?
- ML basic idea
  - Map input data to output data
  - Typical map task: classification
- Using O3:
  - Prepare data in tab or comma (,) file format
  - Import with File widget
  - Review data in Data Table widget
  - Add Data Sampler widget; choose sample
  - The following sample parameters good for first model
  - Add widgets on next slide to make regression model
  - Look at the Confusion Matrix

- Add Test and Score widget
- STUDY RESULTS!

- DEMONSTRATION O3 ML ON "ASSOCIATION RULE MINING"(FOOD DATA)
  - Minimal support: percentage of the entire data set covered by the entire rule (antecedent and consequent)
  - Minimal confidence: proportion of the number of examples which fit the right side (consequent) among those that fit the left side (antecedent)
  - Max. number of rules: limit the number of rules the algorithm generates. Too many rules can slow down the widget considerably.

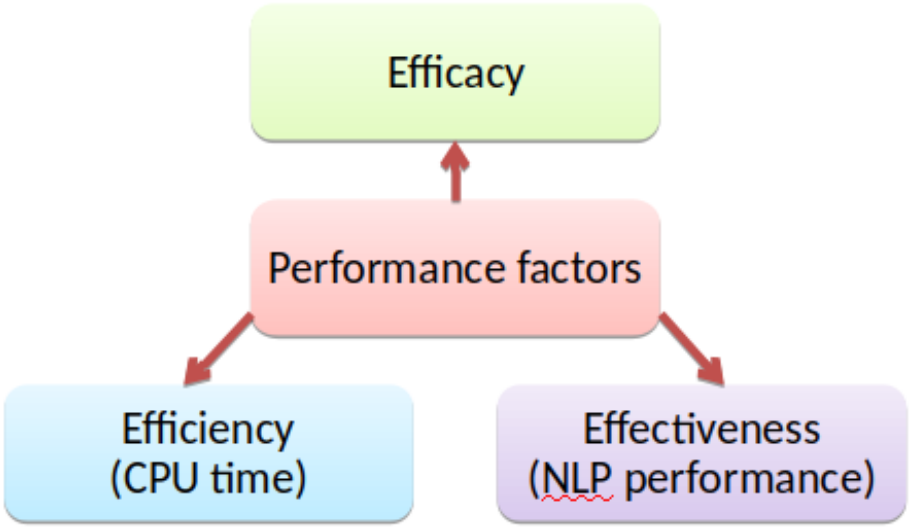
11 WEEK 13: WORD EM BEDDING S IN NLP: WHY AND HOW

- BASIC IDEAS: BAG OF WORDS, TEXT AS DATA
  - BoW ignores document structure (e.g., "POS", section headers...)
  - Input = corpus à Output = one vector per document
  - BoW counts each token/word in clinical note (or sentence)
  - Convert each note à vector ("vector space model")
  - Usually word/token frequencies are normalized ("tf-idf" etc.)
  - BoW learning works best with VERY LARGE corpora
  - No labels required (good)
  - Minimal pre-processing (good)
  - No parsing, no linguistics, no dictionary (good)
  - Problem: corpus 1,000 documents with 10,000 unique words == matrix of 10^7 elements
  - Vectors very sparse (many 0 counts)

- BASIC IDEAS: WORD EMBEDDINGS
  - Word vectors are positioned in the vector space so words that share common contexts in the corpus are located in close proximity in the space.

- "down"dimensionality AND keeps much structure
- Meanings of words (semantics) often retained
- Popular algorithm: “Word2vec” (old)
- Word2vec makes vectors and “learns” vector space
- Word2vec computes word embeddings using “neural network”
- Word embeddings == LARGE corpus == slow
- “Graphical Processing Units” (GPUs) down compute time
- Once computed, word embeddings easily shared
- Vector elements stored as “real numbers”
- Fixed length vectors(<,1000)

12 Week 14, SDLC Testing



- Validation: is the process of measuring efficacy
  - Measures how accurately the NLP system performs extraction or classification as compared to a reference standard
  - Reference standard can be
    - Manually annotated text

    - Benchmark system
  - Measured in classic performance measures
    - Classification accuracy
    - Recall
    - Precision
    - F-measure
  - Error analysis

- Classification
  - assigning a label to an instance (mention, document, patient)
  - $Classification\_accuracy = \frac{Matched}{Total\_number\_classified}$
  - Binary classification = two labels

System	Reference standard		
	Class 1	Class 2	
Class 1	True positive Class 1 Matched	Not matched	Total system Class 1
Class 2	Not matched	True positive Class 2 Matched	Total system Class 2
	Total ref standard Class 1	Total ref standard Class 2	Total number classified

System	Reference standard		
	Class 1	Class 2	
Class 1	15	30	45
Class 2	5	950	955
	20	980	1000

Accuracy = (15+950)/1000 = 0.965

System	Reference standard		
	Class 1	Class 2	
Class 1	0	0	0
Class 2	20	980	1000
	20	980	1000

Accuracy = (0+980)/1000 = 0.980

  - Multi-class classification = three or more labels

System	Reference Standard			
	Class 1	Class 2	Class 3	
Class 1	True positive Class 1 Matched	Not matched	Not matched	Total system Class 1
Class 2	Not matched	True positive Class 2 Matched	Not matched	Total system Class 2
Class 3	Not matched	Not matched	True positive Class 3 Matched	Total system Class 3
	Total ref standard Class 1	Total ref standard Class 2	Total ref standard Class 3	Total number classified

– Confusion Metrics:

* Binary

System	Reference standard		
	Class 1	Class 2	
Class 1	a = 15	c = 30	a + c = 45
Class 2	b = 5	d = 950	b + d = 955
	a + b = 20	c + d = 980	a + b + c + d = 1000

$PPV_{class1} = \frac{a}{a+c} = \frac{15}{45} = 0.33$

$Sensitivity_{class1} = \frac{a}{a+b} = \frac{15}{20} = 0.75$

$NPV_{class1} = \frac{d}{b+d} = \frac{950}{955} = 0.99$

$Specificity_{class1} = \frac{d}{c+d} = \frac{950}{980} = 0.97$

$PPV_{class2} = \frac{d}{d+b} = \frac{950}{955} = 0.99$

$Sensitivity_{class2} = \frac{d}{d+c} = \frac{950}{980} = 0.97$

$NPV_{class2} = \frac{a}{c+a} = \frac{15}{45} = 0.33$

$Specificity_{class2} = \frac{a}{b+a} = \frac{15}{20} = 0.75$

$Accuracy = \frac{a+d}{a+b+c+d} = \frac{950+15}{1000} = 0.965$

$Precision_{class1} = \frac{a}{a+c} = \frac{15}{45} = 0.33$

$Recall_{class1} = \frac{a}{a+b} = \frac{15}{20} = 0.75$

$Precision_{class2} = \frac{d}{d+b} = \frac{950}{955} = 0.99$

$Recall_{class2} = \frac{d}{d+c} = \frac{950}{980} = 0.97$

$Classification\_Error = 1 - Accuracy = 1 - 0.965 = 0.035$

Macro-averaging = calculating simple average over all classes

$Precision_{macro} = (Precision_{class1} + Precision_{class2})/2 = (0.33 + 0.99)/2 = 0.66$

$Recall_{macro} = (Recall_{class1} + Recall_{class2})/2 = (0.75 + 0.97)/2 = 0.86$

$Accuracy_{macro} = 0.97$

Micro-averaging - calculating combined metrics over all classes

$Precision_{micro} = \frac{950+15}{1000} = 0.965$

$Recall_{micro} = \frac{950+15}{1000} = 0.965$

$Accuracy_{micro} = \frac{950+15}{1000} = 0.965$

F1-Score = harmonic mean between precision and recall

$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$

$F1_{class1} = \frac{2 \times 0.33 \times 0.75}{0.33 + 0.75} = 0.459$

$F1_{class2} = \frac{2 \times 0.99 \times 0.97}{0.99 + 0.97} = 0.982$

$F1_{macro} = \frac{2 \times 0.66 \times 0.86}{0.66 + 0.86} = 0.75$

* Multi

– Confusion metrics (contingency table)

System	Reference standard			
	Class 1	Class 2	Class 3	
Class 1	a	d	g	a+d+g
Class 2	b	e	h	b+e+h
Class 3	c	f	j	c+f+j
	a+b+c	d+e+f	g+h+j	a+b+c+d+e+f+g+h+j

$$PPV_{Class1} = \frac{a}{a+d+g} = \frac{TP_{Class1}}{Total\ count\ System\ Class1}$$

$$Sensitivity_{Class1} = \frac{a}{a+b+c} = \frac{TP_{Class1}}{Total\ count\ RefSt\ Class1}$$

– Confusion metrics (contingency table)

System	Reference standard			
	Class 1	Class 2	Class 3	
Class 1	a	d	g	a+d+g
Class 2	b	e	h	b+e+h
Class 3	c	f	j	c+f+j
	a+b+c	d+e+f	g+h+j	a+b+c+d+e+f+g+h+j

$$PPV_{Class1} = \frac{a}{a+d+g} = \frac{TP_{Class1}}{Total\ count\ System\ Class1}$$

$$Sensitivity_{Class1} = \frac{a}{a+b+c} = \frac{TP_{Class1}}{Total\ count\ RefSt\ Class1}$$

$$NPV_{Class1} = \frac{e+h+f+j}{b+c+e+h+f+j} = \frac{TN_{Class1}}{Total\ count\ System\ Not\ Class1}$$

$$Specificity_{Class1} = \frac{e+h+f+j}{d+g+e+h+f+j} = \frac{TN_{Class1}}{Total\ count\ RefSt\ Not\ Class1}$$

06/18/2019 © Patterson 2017-2018 14

- Extraction
  - TP: True Positive => Reference standard and System output match exactly
  - FP: False Positive => System output without exact match to reference standard
  - TN: True Negative => No exact matches
  - FN: False Negative => Reference standard without exact match to system output
  - LTP: Loose True Positive => Reference standard and system output overlap
  - LFP: Loose False Positive => System output does not overlap reference standard
  - LFN: Loose False Negative => Reference standard does not overlap system output

System	Reference standard		
	Class 1	Class 2	
Class 1	TP	FP	Total system positive
Class 2	FN	TN	Total system negative
Class 2	Total ref standard positive	Total ref standard negative	

- Error analysis
  - manually examine the examples when system output did not match manual annotation and attempt to find the reason for the erro
  - Systematic error
    - The same reason for each instance of an error of that type

    - Can be fixed by updating dictionary, or rules, or retraining machine learning step with a different feature set
  - Random errors
    - New misspellings, new context
    - Each error is different from other errors
    - Cannot be fixed without significant decrease in recall or precision

- Effectiveness
  - Refers to the ability of the system to achieve the required goal
  - Most cases it is the same as efficacy
  - For some use cases, extracting some instances with higher precision is more important than extracting all instances.
  - Case1: Finding at least one mention is enough.

	Document 1	Document 2	Document 3	Total
Matches	2	1	0	3
System labeled	3	2	0	5
Ref standard labeled	2	2	2	6
Effective True positive	1	1	0	2

    Mention level Precision = 3 / 5 = 0.60

    Effective Document level Precision = 2 / 2 = 1.00

    Mention level Recall = 3 / 6 = 0.50

    Effective Document level Recall = 2 / 3 = 0.66

  - Case 2: Finding all instances is essential.

	Document 1	Document 2	Document 3	Total
Matches	2	1	0	3
System labeled	3	2	0	5
Ref standard labeled	2	2	2	6
Effective True positive	1	0	0	1

    Mention level Precision = 3 / 5 = 0.60

    Effective Document level Precision = 1 / 2 = 0.50

Mention level Recall = 3 / 6 = 0.50

Effective Document level Recall = 1 / 3 = 0.33

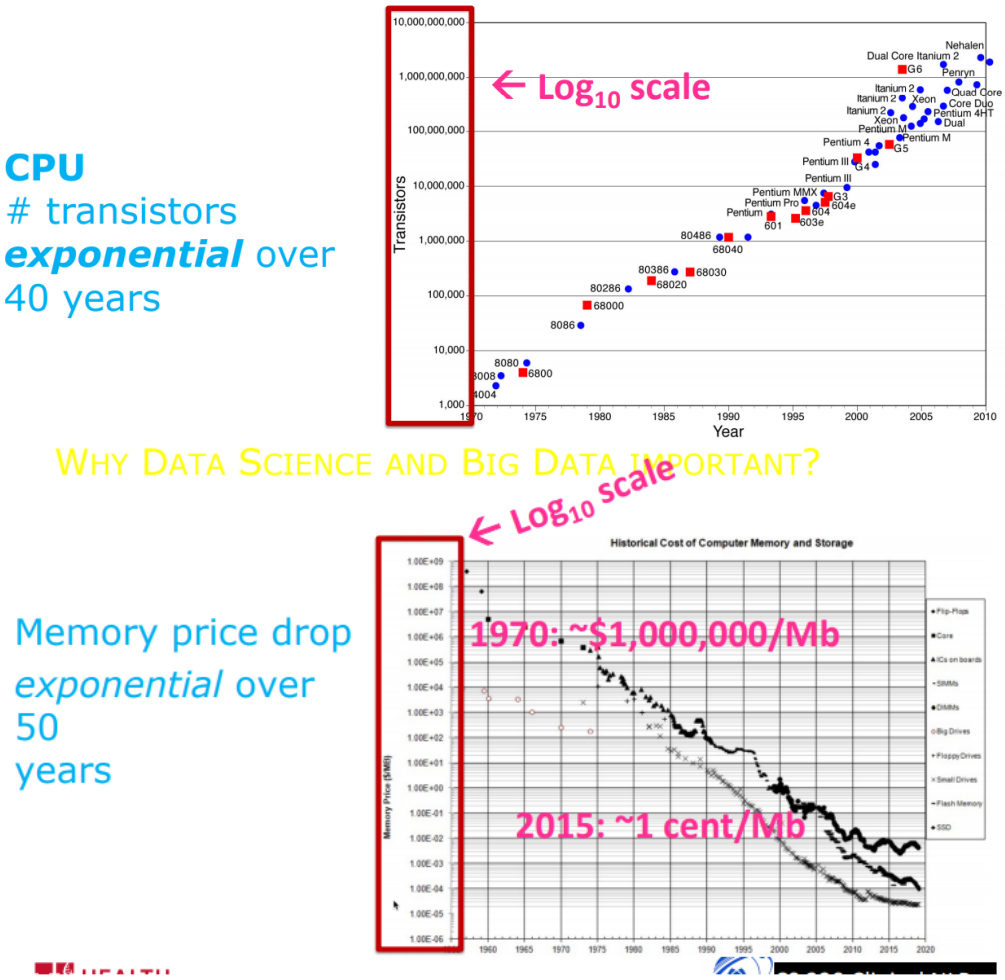
- Efficiency
  - refers to system performance, or computational efficiency
  - measured in documents processed per second
  - Benchmarking and processing optimization is required for larger projects

13 Week 15, Course Review

- DATA SCIENCE SUCCESS: HARDWARE Other "AI" apps show intelligence?
  - Speech recognition?
  - Best player of Go?
  - Best player of chess?
  - Best player of American game show Jeopardy
  - Google/Facebook can predict personal ads (advertisments) for you?
- DATA SCIENCE SUCCESS: HARDWARE
  - Basic ML model theory > 30+ years old
  - Regression statistics
  - Classification Trees (CART Breiman 1984, 2nd gen == Random Forest)
  - Neural Network theory (Back propagation 1986)
  - SVM (Vapnik & Chervonenkis 1963)
  - So why Deep learning now?
  - Market driver behind Deep Learning: speech recognition
    - Needed a real-world use case...
    - Mobile technology (digital phones) in wide use
    - in England (#of digital calls) > (# hardwired)

    calls in 2011.

- some pictures



WHY DATA SCIENCE AND BIG DATA IMPORTANT?

- DATA SCIENCE APPS US HEALTHCARE: SOTA
  - Python is the data language for data science
  - Scikit-learn most popular Python ML toolset
  - Orange3 embeds many scikit-learn tools
  - Jupyter notebooks becoming the standard for development and research

14 QUIZES

14.1 QUIZ 1

1. Question 1

   Q: In the article "Building Resources Vietnamese Clinical Text Processing"the authors annotated [x]

and had high precision.

A: clinical not

2. Question 2

   Q1: The collection of the Health Ministry, all hospitals, and all clinics is called

   A1: The Healthcare system

   Q2: A health NLP system finds a "mention"(or target) when it finds

   A2: a phrase or term

   Q3: A health NLP system can "assert"a classification at the level of a mention, a note, or ...

   A3: an encounter

3. Question 3

   Q: True or false: Health NLP is used NOW in some cell phone apps (USA) or used by the servers that support cell phone apps (USA).

   A: True

4. Question 4

   Q: At least half of Vietnamese adults have used a mobile (cell) phone?

   A: True

5. Question 5

   Q: Name another source of text that might be used for health NLP in USA.

   A: news

6. Question 6

   Q: Which app in Anaconda Console is designed for data mining by non-experts?

   A: Orange3

7. Question 7

   Q: Name one source of text that might be used for health NLP in Vietnam.

   A: news

14.2 QUIZ 2

1. Question 1

   Q1: Why is it important to do Analysis as part of the system design (SDLC)?

   A1: Because in SDLC Analysis follows Planning, and keeping order is important

   Q2: Why is it important to do Planning as part of the system design (SDLC)?

   A2: Because some projects may not ever work (not feasible) – important to know before starting.

2. Question 2

   Q: True or false: To a clinical NLP system, we compare human annotations (reference) to the NLP system annotations (machine), and testing is always required.

   A: True

14.3 QUIZ 4

1. Question 1

   Q: A pipeline framework can also be used for non-NLP processing task (i.e., for Applications other than NLP)

   A: True

2. Question 2

   Q: Pipeplines components use input from the previous component's output and provide input to the following component. The first component in a pipeline has no input. That is, if starts the pipelines running.

   A: False. The input to the first component of a pipeline are the data that is being processed by the

pipeline. In the case of health NLP, that input is a text corpus.

3. Question 3

   Q: In the first two weeks of the course we discussed the **context** of mention (A **mention** is an NLP target term). We named threee common types of context in health NLP (historical versus current/active, positive or negative, and current patient versus someone else like a relative). The **Assertion_Classifier** in the default NER pipeline identifies one type of context. It identifies:

   A: Whether the mention is positive or negative assertion (e.g, an example of a negative assertion: The patient denied having a cough).

4. Question 4

   Q: Most projects that use the output of CLAMP process file in the output directory in the current CLAMP project. Which type of output file would be easier to parse?

   A: xxx.txt

5. Question 5

   Q: The default CLAMP-NER pipeline was provided with the CLAMP installation. Another pipelines is provided that extracts medicantion attributes.

   A: True

14.4 QUIZ 5

1. Question 1

   : Q: True or false: in CLAMP mapping terms to categories happens once in a pipeline and categories can be used in an NLP rule or query by another pipeline stage.

   : A: True, Once annotated by CLAMP, categories can be used in any "downstream"pipeline stage.

2. Question 2

   Q: Not true about O3.

   A: Orange3 runs best on its own download data., Orange3 does not allow users to insert their own

Python code in a widget.

3. Question 3

   Q: CLAMP uses dictionaries and parts-of-speech (POS) in many pipelines. POS processing is required in any clinical NLP system.

   A: False.

4. Question 4

   Q: In the "SMOKER" example above, three terms are mapped to one category. Pick the best statement about categories: A: Categories are more general than specific terms and can simplify writing an NLP rule.

14.5 QUIZ 6

1. Question 1

   A: Vietnam has less [female obesity] than Australia. But Vietnam has about the same [female diabetes] as Australia.

2. Question 2

   A: The American agency [NIH] created [UMLS], a collection of over 100 different sources of healthcare terms and definitions.

3. Question 3

   Q: Select the two best statements about the CLAMP clinical NLP system.

   A: The correct answer is: CLAMP pipelines usually start with a sentence segmenter., CLAMP allows users to write their own RUTA rules in some pipeline stages.

4. Question 4

Q: To increase the data set size for Orange3, we added notes from a Harvard corpus. We had to process those notes offline because sharing the Harvard files with students is not allowed. True of false: if CLAMP takes N seconds to process 200 default corpus notes, then we expect CLAMP takes N seconds to process 200 Harvard notes.

A: The Harvard and default note size might be very different. In fact, Harvard notes are far larger than default notes on average. CLAMP takes more time processing Harvard notes. The correct answer is 'False'.

14.6 QUIZ 7

1. Question 1

   A: Orange3 is not a perfect machine learning (ML) tool, but it has several good features. One good feature is that every ML project needs a [baseline] and O3 builds that quickly. A second good feature is that O3 helps find [data problems] early in ML design.

2. Question 2

   Q: True or false: machine learning is basically mapping a set of inputs to a set of outputs. A: True

3. Question 3

   Q: Pick the one answer that is not correct about machine learning and NLP.

   A: If an ML model uses data from an NLP system as input, then testing usually takes longer (more processing time) than training. ĐúngCorrect, this is the one incorrect statement. First, there is nothing special about NLP output as ML input. Second, testing always takes less processing time than training. In training, the ML has to find the model solution (it "explores"the solution space). In testing, the solved model simple "runs".

4. Question 4

   A: Orange3 is not a perfect machine learning (ML) tool, but it has several good features. One good feature is that team members using it [communicate better]. A second good feature is that less programming means [faster start].