KNOWLEDGE DISCOVERY CHALLENGES ON BIG MEDICAL DATA

Tutorial at ECML PKDD 2014

Nancy, France, 09/2014

Ernestina Menasalvas: Big Medical Data Mining

- PhD (1998) on data mining
- Chair of the MIDAS (Data Mining and Simulation) research group at UPM
- Joined CTB 3 years ago
- Emphasis on text and image processing from EHR
- Strong links with public Hospitals from Madrid

POLITÉCNICA
"Ingeniamos el futuro"
CAMPUS DE EXCELENCIA INTERNACIONAL
CTB center for biomedical technology

Agenda PART III

- Motivation
- EHR
- BIG DATA in the health domain
  - Applications
  - Goal
  - Process
  - Non structured data
- Image Processing
- Text Processing
- Final Thoughts and Conclusions

MOTIVATION

Motivation

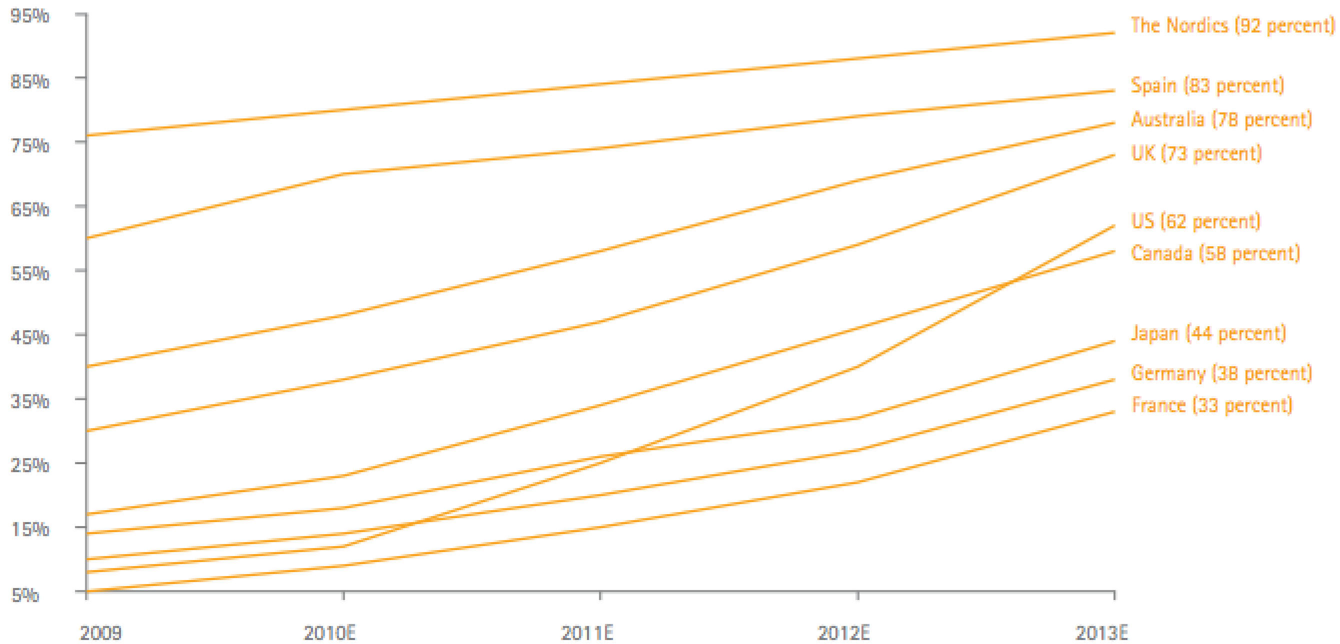
- In 2012, worldwide digital healthcare data was estimated to be equal to 500 petabytes and is expected to reach 25,000 petabytes in 2020
- Can we learn from the past to become better in the future?
- Healthcare Data is becoming more complex !!
- The problem :
  - *Milllions of reports, tasks, incidents, events, images, …*
  - *Complete availability*
  - *Lack of protocols and structure*
  - *Organization oriented processes*
- Need of patient oriented processes → information

From Mckensey: big data in health report 2013

- From physicians judgment to evidence-based medicine
- Standard medical practice is moving from relatively ad-hoc and subjective decision making to evidence-based healthcare
- Is the health-care industry prepared to capture big data’s full potential, or are there roadblocks that will hamper its use?
- Holistic, **patient-centered** approach to value, one that focuses equally on health-care spending and treatment outcomes.

ELECTRONIC HEALTH RECORDS (EHR)

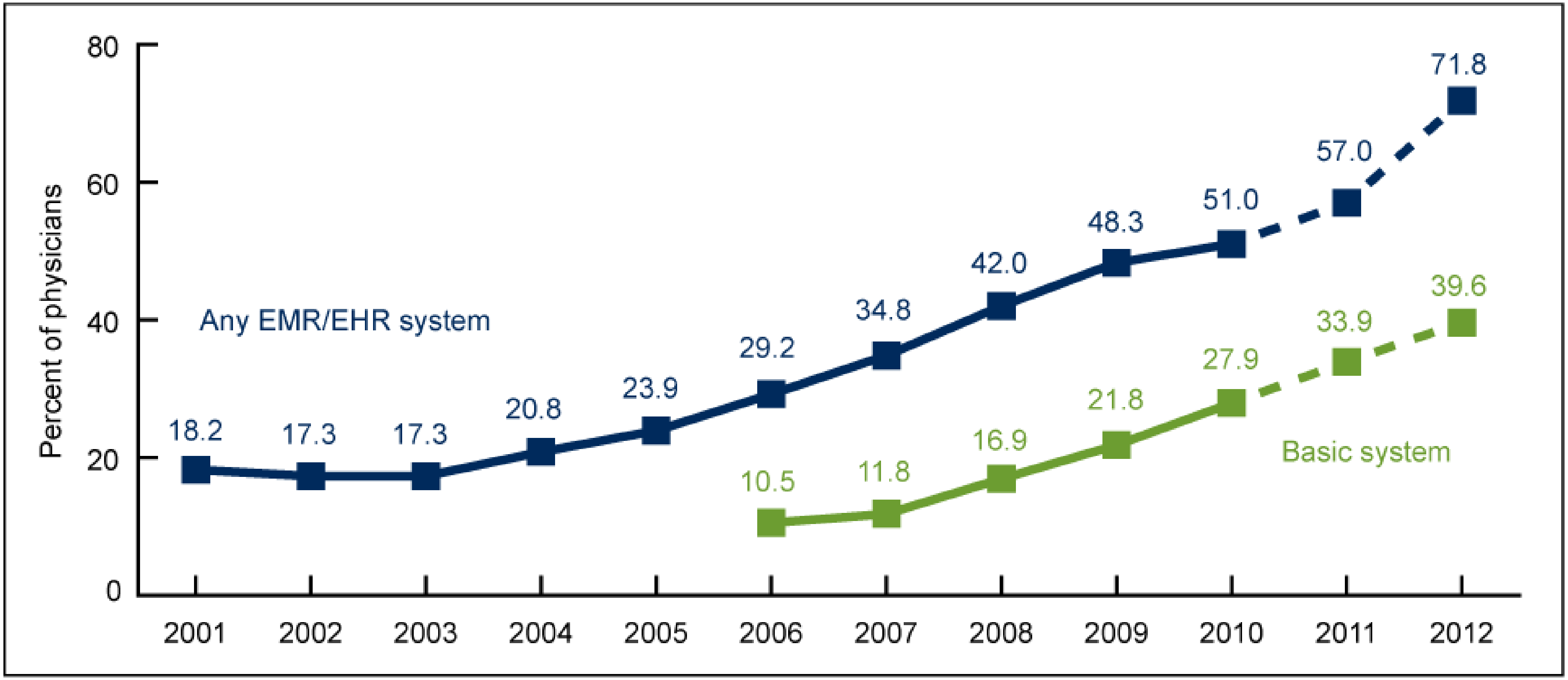
EHR adoption



http://www.accenture.com/SiteCollectionDocuments/PDF/Accenture_EMR_Markets_Whitepaper_vfinal.pdf

EHR adoption

Figure 1. Percentage of office-based physicians with EMR/EHR systems: United States, 2001–2010 and preliminary 2011–2012



NOTES: EMR/EHR is electronic medical record/electronic health record. "Any EMR/EHR system" is a medical or health record system that is all or partially electronic (excluding systems solely for billing). Data for 2001–2007 are from in-person National Ambulatory Medical Care Survey (NAMCS) interviews. Data for 2008–2010 are from combined files (in-person NAMCS and mail survey). Data for 2011–2012 are preliminary estimates (dashed lines) based on the mail survey only. Estimates of basic systems prior to 2006 could not be computed because some items were not collected in the survey. Data include nonfederal office-based physicians and exclude radiologists, anesthesiologists, and pathologists.
SOURCE: CDC/NCHS, National Ambulatory Medical Care Survey, 2001–2012.

EHR Knowledge Extraction

- Electronic Health Records’ use has been increasing in the last ten years.
- Digitalization of patients’ histories have led to enormous data stores.
- Most hospitals do not take advantage of analytic processes to improve patient care.

BIG DATA IN THE HEALTH DOMAIN

The average hospital (300 beds)

- 500.000 patients (reference population)
- 1300 users (250 physicians, 900 nurses and technicsian, 150 administrative tasks)
- Monthly activity:
  - 20.000 consultations, 1300 admissions, 800 interventions 10.000 emergencies
  - 75.000 annotations
  - 25.000 reports
  - 90.000 interdepartamental orders
  - 450.000 lab results (analytical)
  - 13.000 images analysis
  - 24.000 pharmacological prescriptions

Hospital Management

- They require of solutions for
  - cost-reduction policies.
  - efficiency procedures.
  - establishing share-risk policies
  - Alarms
  - Early prognosis and diagnosis
  - Environmental, sensor, … integration
  - Use data and services of the cloud for comparison of data of other hospitals/countries/.. for efficiency policies.
  - ..

Goverment

- support for cost-reduction policies
  - analysis of early detection of chronic diseases
  - analysis of diseases and the elderly
  - prediction of the evolution of diseases depending on clinical and societal factors
  - ….
- sentiment analysis (user satisfaction) of policies, health care, …
- impact of environmental factors on the evolution, prevalence and .. of diseases
- impact of socio economic situation of people on the disease evolution and impact on health costs
- cloud based services for analysis of all the data generated in different hospitals

Clinicians: evidence based medicine

- correlations, associations of symptoms, familiar antecedents, habits, diseases
- impact of certain biomedical factors (genome structure, clinical variables) on the evolution of certain diseases
- automatic classification of images (prioritization of RX images to help diagnosis)
- automatic annotation of images
- natural language (google style) based diagnose aid tools

Researchers

- find early indicators of diseases
- design of clinical trials
- automatic search in bibliography using not only keywords but also analyzing the text of the papers
- use of analytics services available on the web
- Use data and services of the cloud for in order to obtain knowledge from of other hospitals/countries/...

Goal

Provide **right** intervention to the **right** patient at the **right** time

ACQUIRE,

PROCESS,

ANALYZE

UNDERSTAND

↓

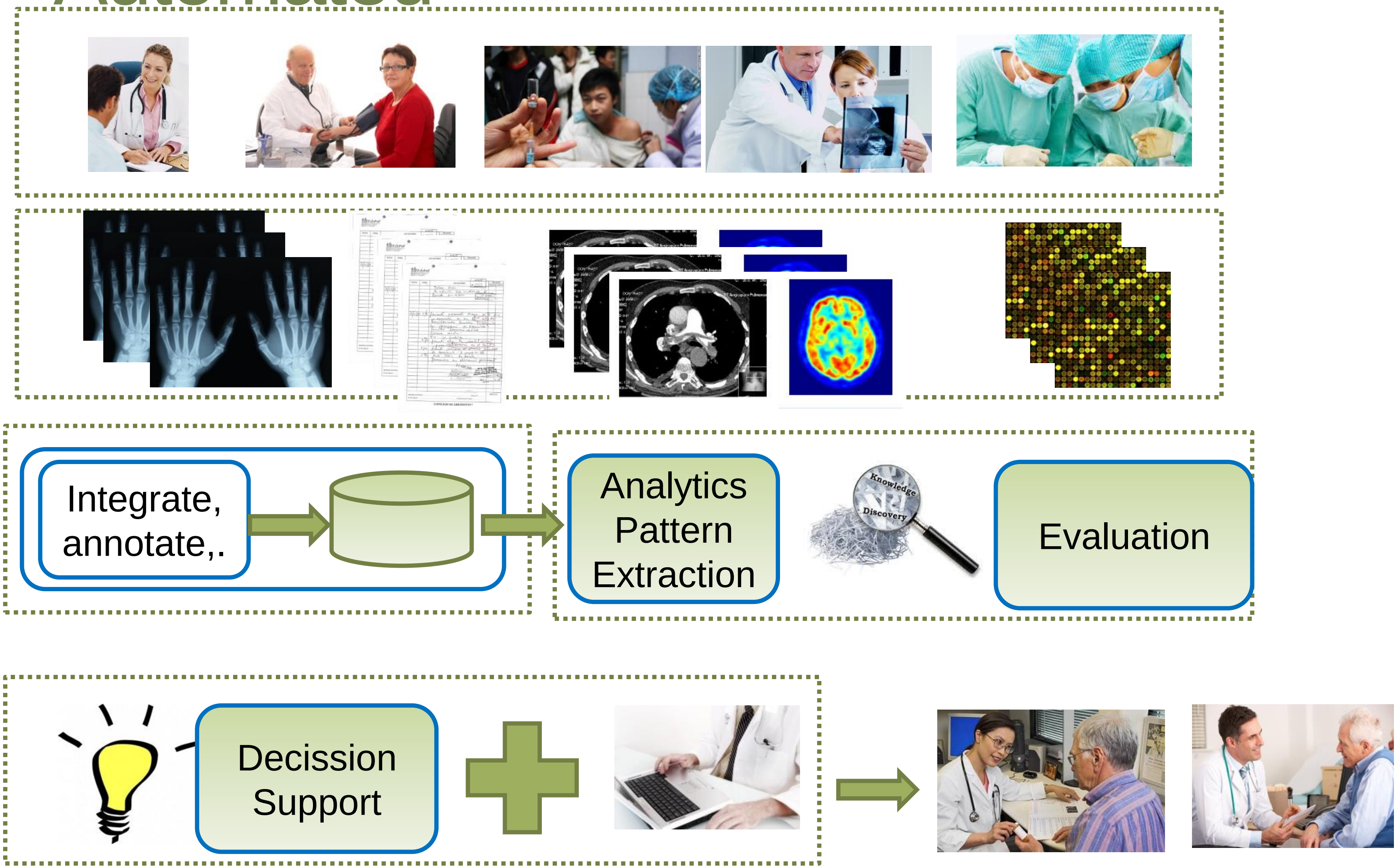
PREDICT

Goal

- Prediction will enable
  - Personalized care to the patient.
  - Early diagnose
  - Lower cost
  - Improved outcomes
  - …

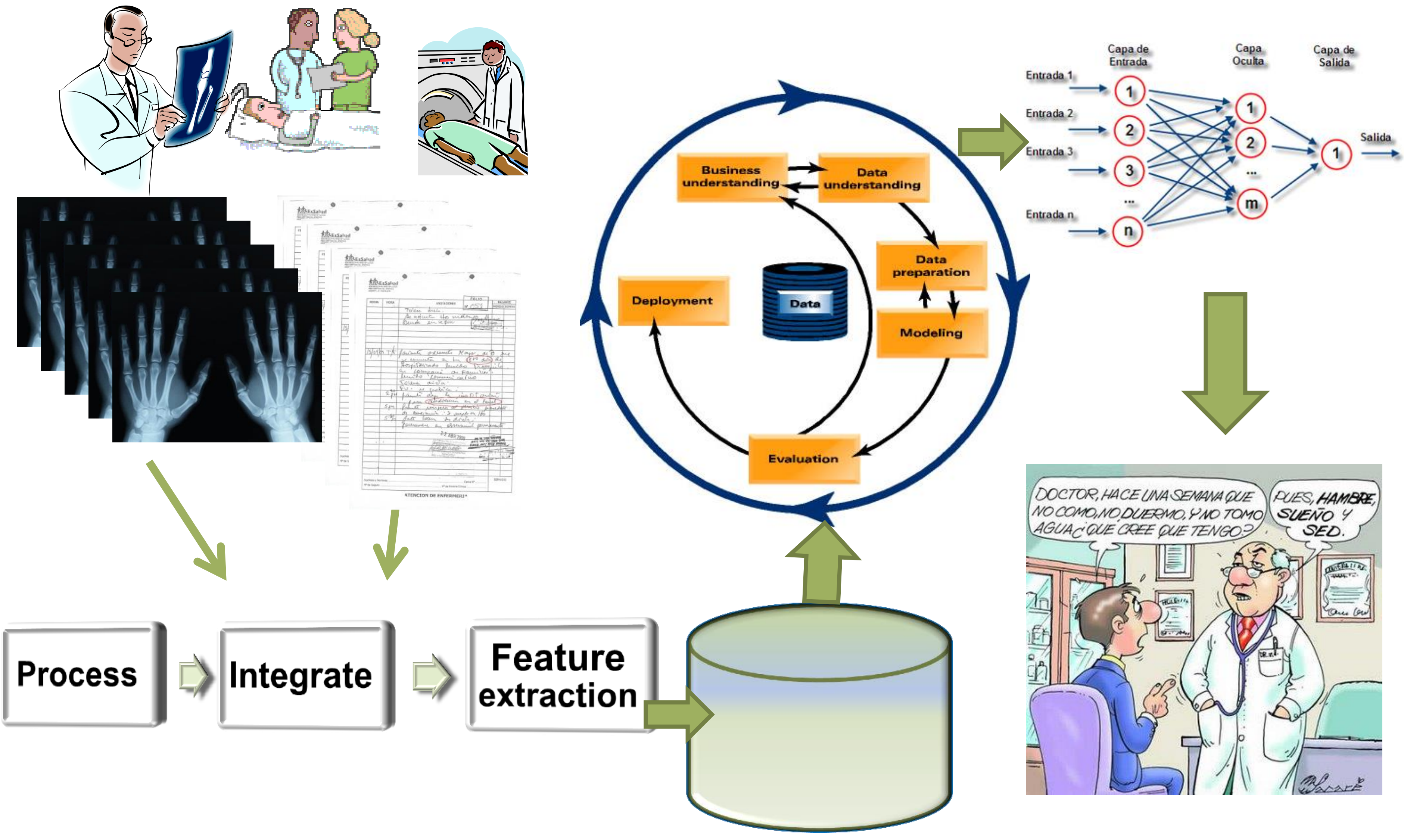
Traditionally

Automated



Integrate, annotate,.
Analytics Pattern Extraction
Knowledge
Discovery
Evaluation
Decission Support

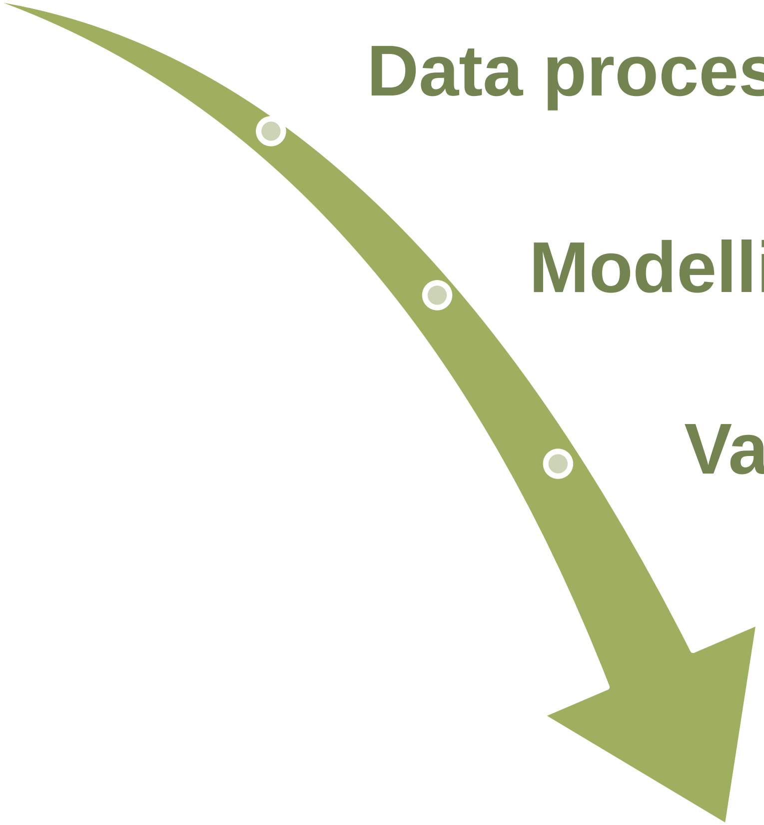
Process



Business understanding
Data understanding
Data preparation
Modeling
Evaluation
Deployment
Data
Capa de Entrada
Capa Oculta
Capa de Salida
Entrada 1
Entrada 2
Entrada 3
Entrada n
Salida
DOCTOR, HACE UNA SEMANA QUE NO COMO, NO DUERMO, Y NO TOMO AGUA ¿QUÉ CREE QUE TENGO?
PUES, HAMBRE, SUEÑO Y SED.
Process
Integrate
Feature extraction

Process

Data Acquisition



Data processing

Modelling

Validation

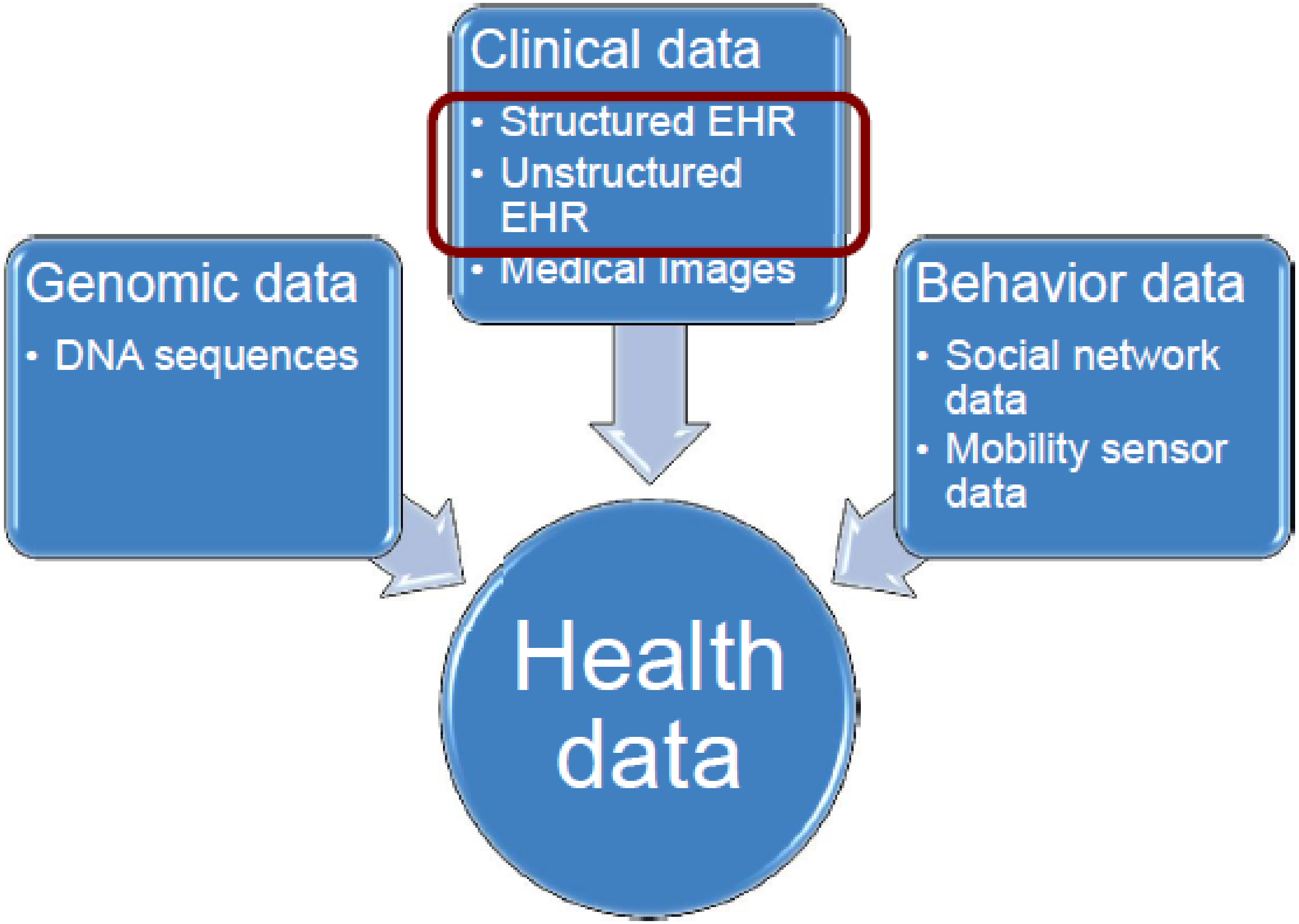
Apply

1st step: Data acquisition

- EHR:
  - Structured data:
    - Lab tests (LOINCR)
      - Many lab systems still use local dictionaries to encode labs
      - Diverse numeric scales on different labs
      - Missing data
    - Clinical and demographic data (ICD): 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)
      - Pros: Universally available
      - Cons: medium recall and medium precision for characterizing patients
  - Non-structured data:
    - **Images**
    - **Clinical notes**

2nd step: analysis of the data

- Image annotation
- Natural language processing
- Integration


Clinical data
• Structured EHR
• Unstructured EHR
• Medical images
Genomic data
• DNA sequences
Behavior data
• Social network data
• Mobility sensor data
Health data

Standards

- MeSH (Medical Subject Headings) - A thesaurus for indexing articles for PubMed.
- UMLS (Unified Medical Language System) - Integrates key terminology among different coding standards.
- SNOMED CT - Standard for clinical terminology.
- DICOM (Digital Imaging and Communications in Medicine) - Standard for processing medical images.
- GS1 standards - Used to identify uniquely different medical products.
- LOINC (Logical Observation Identifiers Names and Codes) - Standard for identifying laboratory and clinical observations.
- RxNORM - Standard normalizing names for pharmacy & drugs products.

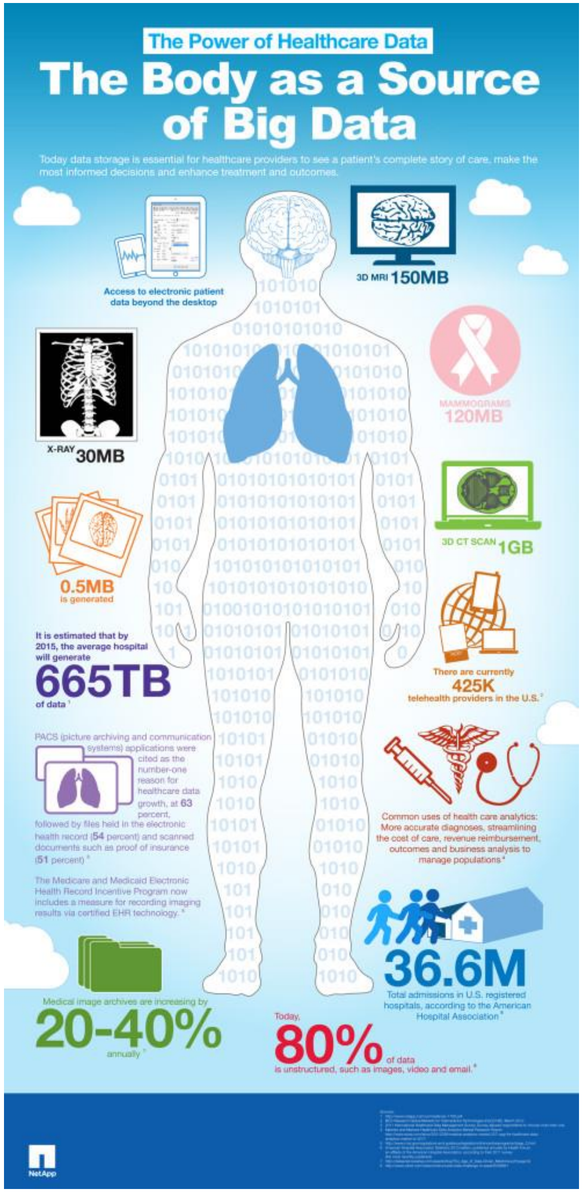
Other resources

- SEDOM provides on its webpage an abbreviations dictionary with 4368 Spanish acronyms.
- Medilexicon (http://www.medilexicon.com) - provides more than 200,000 acronyms.
- OBO Foundry (http://www.obofoundry.org) - provides several biological and biomedical ontologies.
- As well as BFO (http://www.ifomis.org/bfo) - which provides basic ontologies.
- CIMI (http://informatics.mayo.edu/CIMI/index.php/Main_Page) From Mayo clinic provides a Modeling initiative.

MEDICAL IMAGE DATA

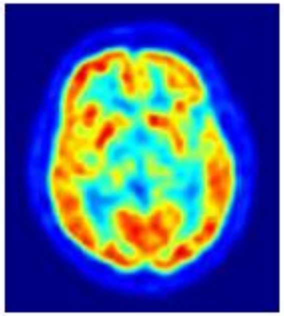
- By 2015, the average hospital will have two-thirds of a *petabyte* of patient data, 80% of which will be unstructured image data like CT scans and X-rays.

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



Most frequent

- **ComputedTomography (CT), X-Ray, Positron Emission Tomography (PET)**
- 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.



PET/CT

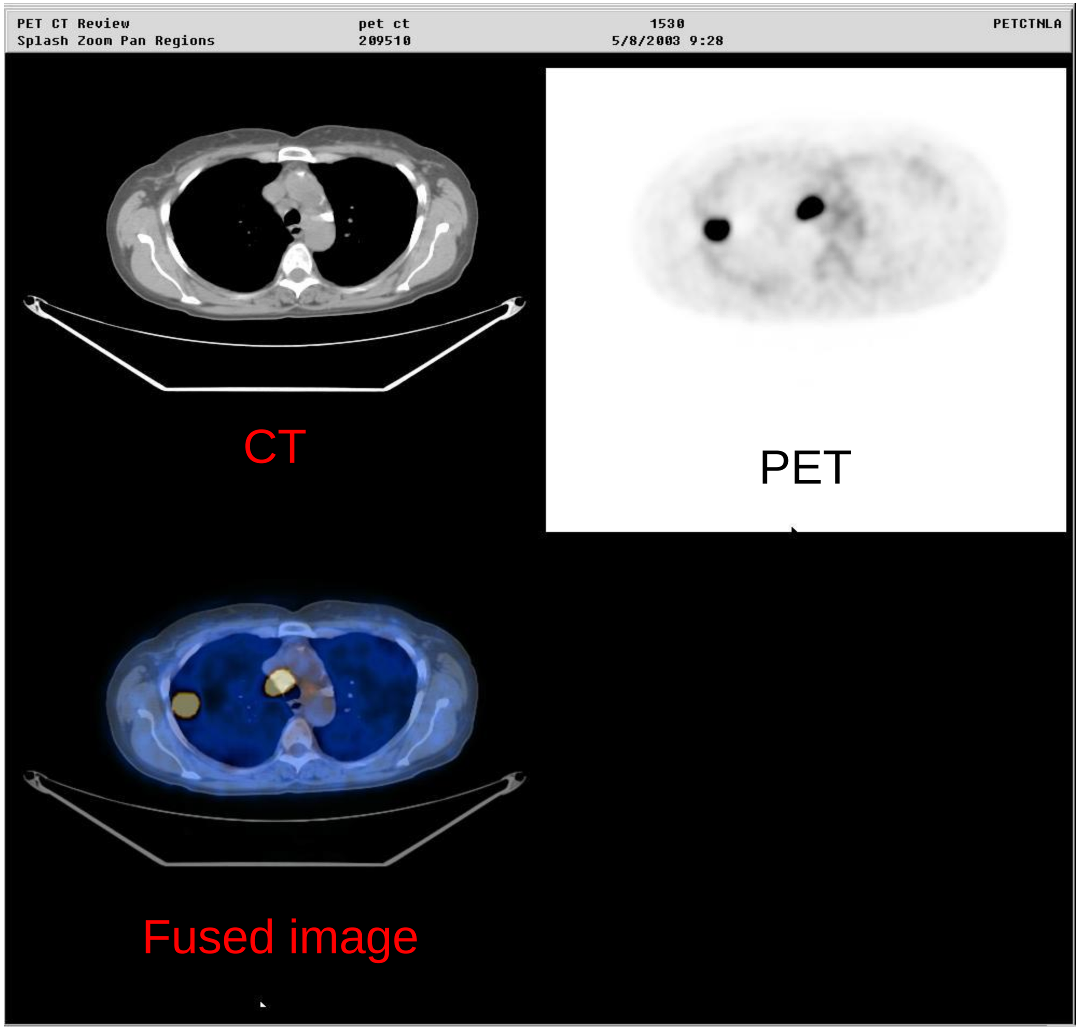
- **Positron Emission Tomography** (PET) and Helical CT
- PET detects area of increased **metabolic activity** as indicated by uptake of radioactive glucose (tumor, infection)
- PET data is usually "fused" with CT data to produce an image showing increased **glucose uptake** superimposed upon the exquisite anatomic detail of helical CT
- Some example of cancers evaluated with PET:
  - Lung
  - Lymphoma
  - Melanoma
  - Colorectal
  - Breast
  - Esophagus
  - Head and Neck

Technical Challenges

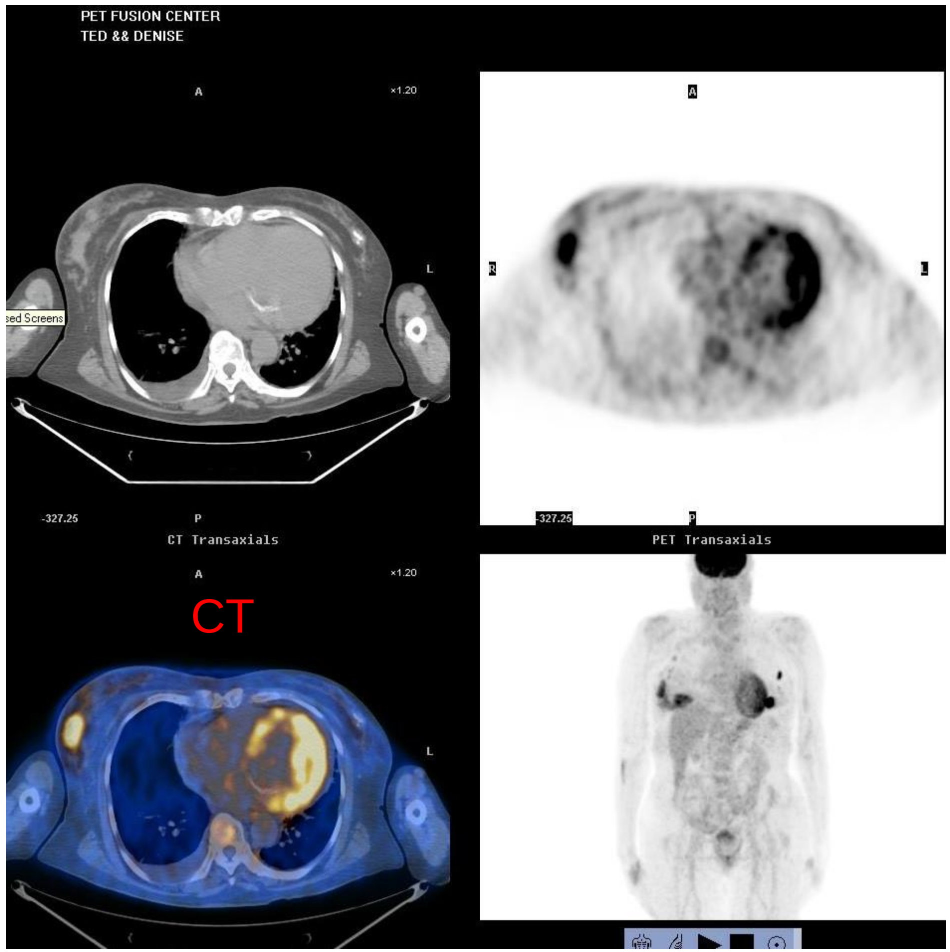
- Imaging Physics - better images by
  - Detector design
    - Spatial resolution
    - Sensitivity
- Radiochemistry - better tracers (PET imaging)

  - **Image processing**
    - **Corrections for physical effects**
    - **Multimodal image fusion**
    - **Image reconstruction algorithms**
- **Data Analysis → better interpretation of images**

Lung carcinoma

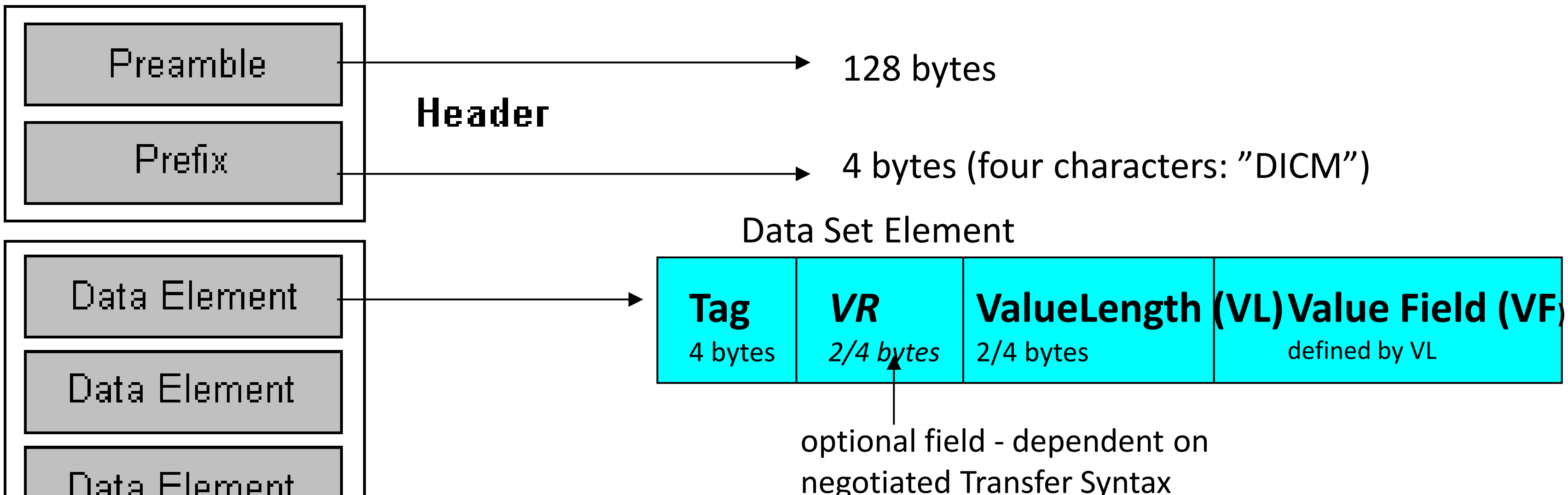


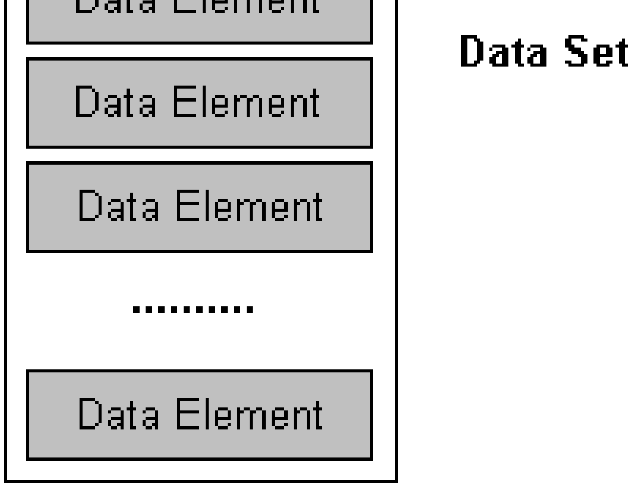
Breast carcinoma



Fused image PET

Metadata Structure





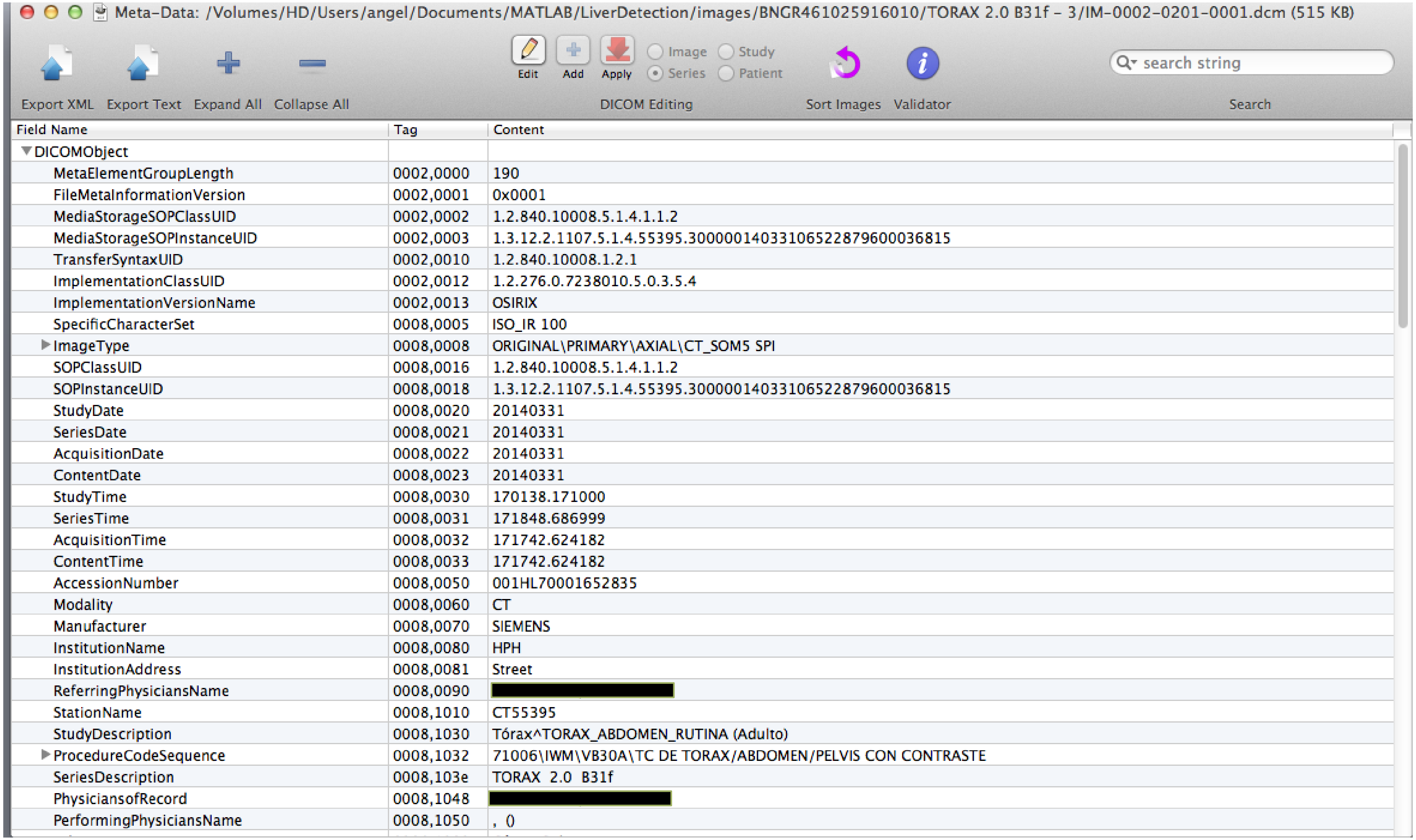
Data Element Tag: 4 bytes, hexadecimal, group (2 bytes), elements(2 bytes)

Value Represention (VR): optional, 2 or 4 bytes, character, depends on transfer syntax

Value Length (VL): 2or 4 bytes, long integer, defines value field length, dependent on VR

Value Field (VF): length is defined by VL, format is defined explicitly by VR or implicitly

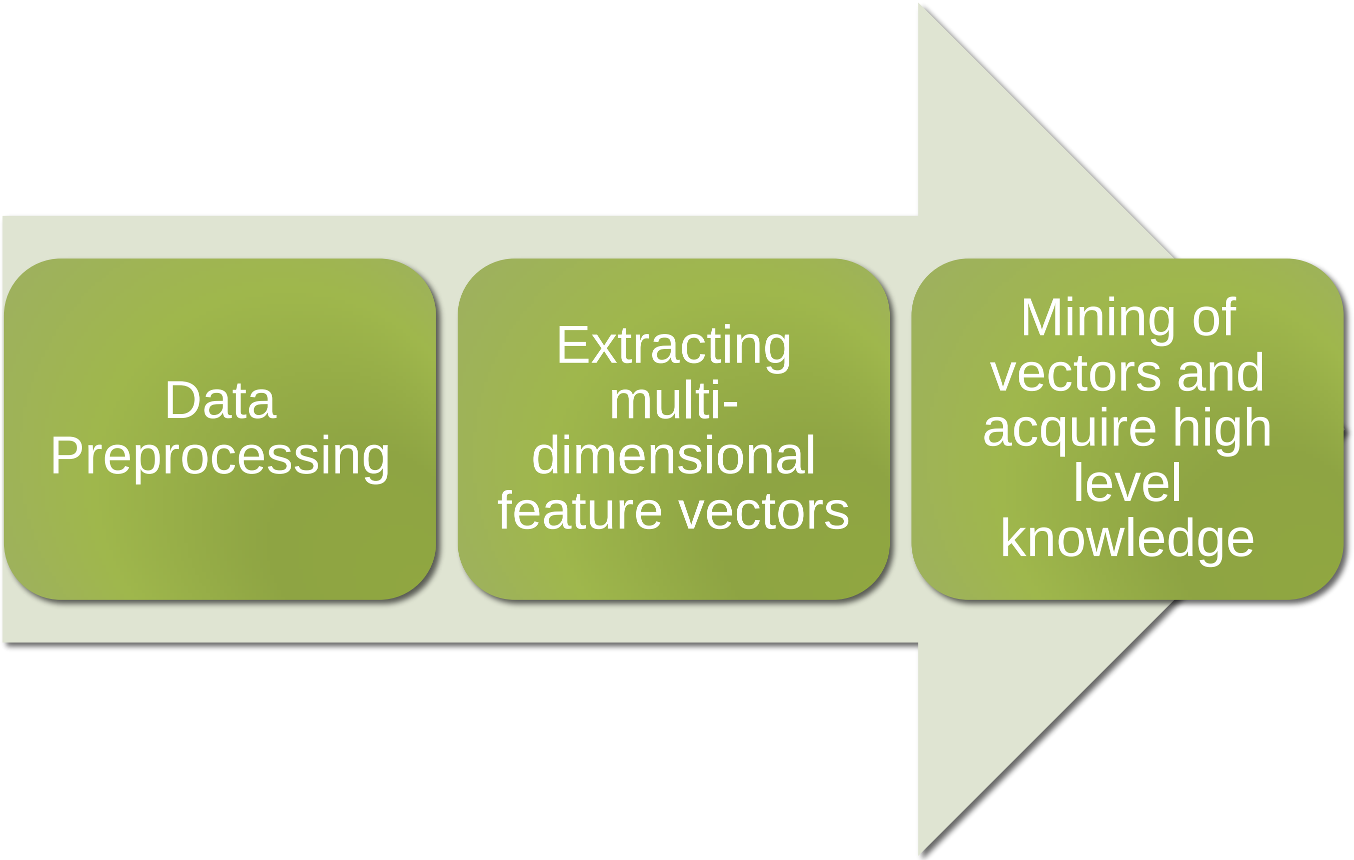
Metadata Structure



Field Name	Tag	Content
DICOMObject		
MetaElementGroupLength	0002,0000	190
FileMetaInformationVersion	0002,0001	0x0001
MediaStorageSOPClassUID	0002,0002	1.2.840.10008.5.1.4.1.1.2
MediaStorageSOPInstanceUID	0002,0003	1.3.12.2.1107.5.1.4.55395.30000014033106522879600036815
TransferSyntaxUID	0002,0010	1.2.840.10008.1.2.1
ImplementationClassUID	0002,0012	1.2.276.0.7238010.5.0.3.5.4
ImplementationVersionName	0002,0013	OSIRIX
SpecificCharacterSet	0008,0005	ISO_IR 100
ImageType	0008,0008	ORIGINAL\PRIMARY\AXIAL\CT_SOM5 SPI
SOPClassUID	0008,0016	1.2.840.10008.5.1.4.1.1.2
SOPInstanceUID	0008,0018	1.3.12.2.1107.5.1.4.55395.30000014033106522879600036815
StudyDate	0008,0020	20140331
SeriesDate	0008,0021	20140331
AcquisitionDate	0008,0022	20140331
ContentDate	0008,0023	20140331
StudyTime	0008,0030	170138.171000
SeriesTime	0008,0031	171848.686999
AcquisitionTime	0008,0032	171742.624182
ContentTime	0008,0033	171742.624182
AccessionNumber	0008,0050	001HL70001652835
Modality	0008,0060	CT
Manufacturer	0008,0070	SIEMENS
InstitutionName	0008,0080	HPH
InstitutionAddress	0008,0081	Street
ReferringPhysiciansName	0008,0090	[redacted]
StationName	0008,1010	CT55395
StudyDescription	0008,1030	Tórax^TORAX_ABDOMEN_RUTINA (Adulto)
ProcedureCodeSequence	0008,1032	71006\IWM\VB30A\TC DE TORAX/ABDOMEN/PELVIS CON CONTRASTE
SeriesDescription	0008,103e	TORAX 2.0 B31f
PhysiciansofRecord	0008,1048	[redacted]
PerformingPhysiciansName	0008,1050	, ()

Methodology for image processing

- Overall process of image mining



Methodology for image processing

1. Data pre-process
   - Calibration: (depending on the device registering the image)
   - Clean up the noise. (noisy pixels)
   - Registration (check the stack of images)
2. Extracting multi-dimensional feature vectors
   - Segmentation Algorithm. Search for homogenous voxels
   - Super-Voxels have to be characterized using low-level features selection
     - Spectral→ digital levels
     - Shape →compactness,..
     - Textural→ smooth, …
     - Context→ neighborhood supervoxels
     - Spatial relationship → up/down,left/right

Methodology for image processing

3. Mining of vectors and acquire high level knowledge
   - Image annotation
   - Indexing and retrieval

Methodology: Image annotation

Image annotation

- Classical approach → manual annotation. It is impractical to annotate a huge amount of images manually
- Second approach → content based image retrieval (CBIR), where images are automatically indexed and retrieved with low level content features like color, shape and texture
- Third approach of image retrieval is the **automatic image annotation**

Methodology: Image annotation

Automatic image annotation

- Single labelling annotation using conventional classification methods: methods (support vector machines (SVM), Artificial Neural Networks, Decision Tree)
  - There three types of AIA approaches:
    - Single labelling annotation using conventional classification methods (support vector machines (SVM), artificial neural network (ANN), and decision tree (DT))
    - Multi-labelling annotation → annotates an image with multiple concepts using the Bayesian methods

Methodology: Image annotation

Automatic image annotation

- Single labelling annotation using conventional classification methods: methods (support vector machines (SVM), Artificial Neural Networks, Decision Tree)
- Binary classification → Tumor / non-tumor cell

Methodology: Image annotation

Automatic image annotation

- Multi-labelling annotation → annotates an image with multiple semantic concepts/categories using the Bayesian methods
- Concept of multi-instance multi-label (MIML) represents an image with a bag of features or a bag of regions. The image is annotated with a concept label if any of the regions/instances in the bag is associated with the label. Then the image is annotated with multiple labels

Methodology: Image annotation

Multi-labelling annotation

- Given a set of images $\{I_1, I_2, \ldots, I_N\}$ from a set of given semantic classes $\{c_1, c_2, \ldots, c_n\}$

Bayesian models try to determine the posterior probability from the priors and conditional probabilities

- Model of conditional approaches
  - Non-parametric
  - Parametric

Methodology: Image annotation

Model of conditional approaches

- Non-parametric. No prior assumption about the distribution of the image features is considered. The actual feature distribution is learned from the features of the training samples using certain statistics.

Automated image annotation: non-parametric approach

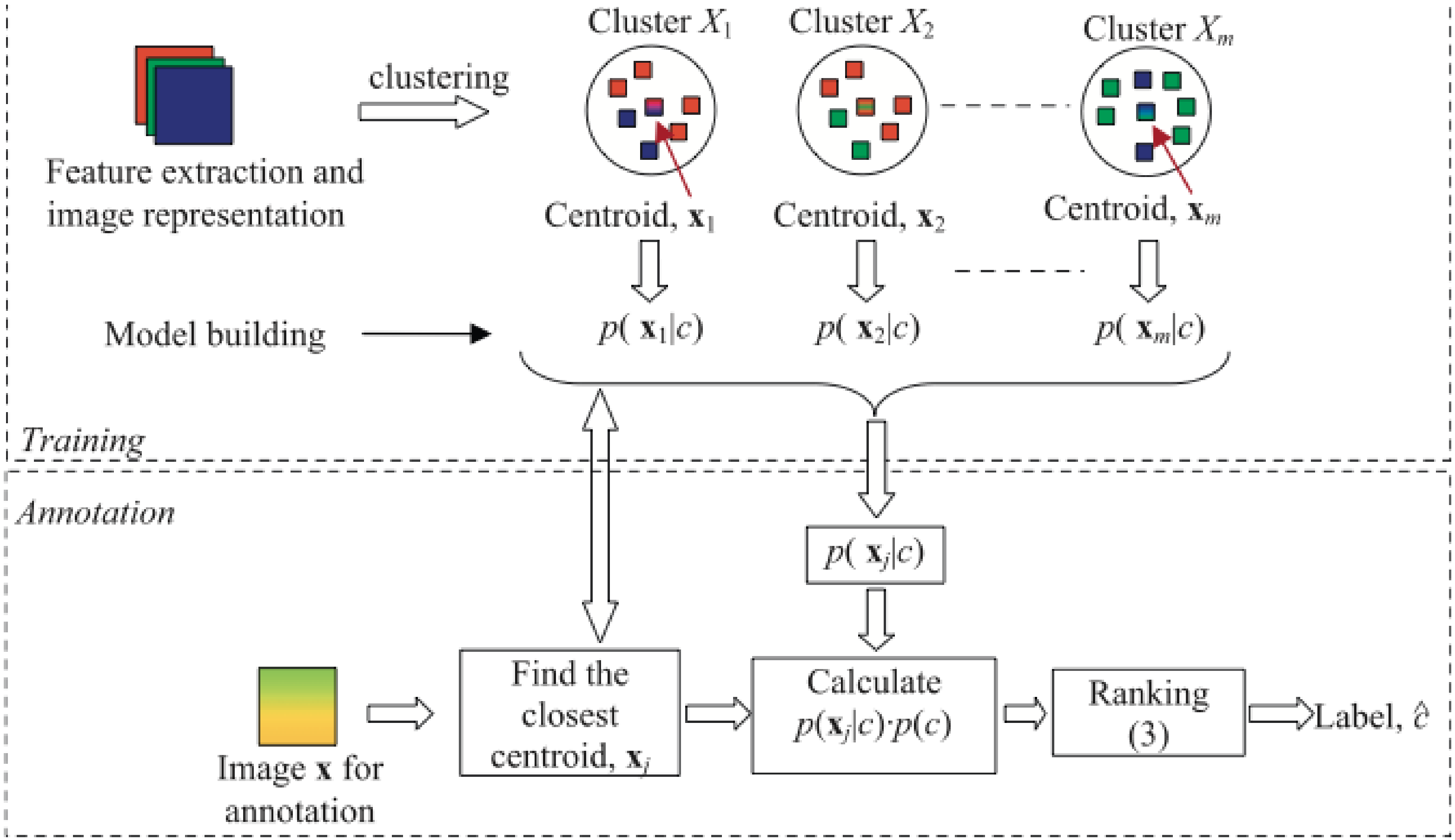


Fig. General Bayesian annotation model

Source: D. Zhang, M. M. Islam, and G. Lu, "A review on automatic image annotation techniques," Pattern Recognition, vol. 45, no. 1, pp. 346–362, Jan. 2012.

Methodology: Image annotation

Model of conditional approaches

- Parametric. The feature space is assumed to follow a certain type of known continuous distribution. Therefore, the conditional probability is modelled using this feature distribution and it is usually modelled as a multivariate Gaussian distribution.

Methodology: Image annotation

Contrast of different annotation methods.

Annotation method	Pros	Cons
SVM	Small sample, optimal class boundary, non-linear classification	Single labelling, one class per time, expensive trial and run, sensitive to noisy data, prone to over-fitting
ANN	Multiclass outputs, non-linear classification, robust to noisy data, suitable for complex problem	Single labelling, sub-optimal, expensive training, complex and black box classification
DT	Intuitive, semantic rules, multiclass outputs, fast, allow missing values, handle both categorical and numerical values	Single labelling, sub-optimal, need pruning, can be unstable
Non-parametric	Multi-labelling, model free, fast	Large number of parameters, large sample, sensitive to noisy data
Parametric	Multi-labelling, small sample, good approximation of unknown distribution	Predefined distribution, expensive training, approximated boundary

Methodology: Indexing and retrieval

Indexing and retrieval

- Two different frameworks
  - Text-based
  - Content-based
- Researh areas
  - Low-level image feature extraction
  - Similarity measurement
  - Deriving high level sematic features

Methodology: Indexing and retrieval

- Levels of queries in CBIR:

  - Level 1: retrieval by primitive features (color, texture, spatial location,…).- Eg.: “find pictures like this”
  - Level 2: retrieval of objects of given type identified by derived features, with some degree of logical inference.- Eg.:”find a picture of a flower”
  - Level 3: retrieval by abstract attributes (emotional, religious,.., significance). Eg.: “find pictures of a joyful crowd”

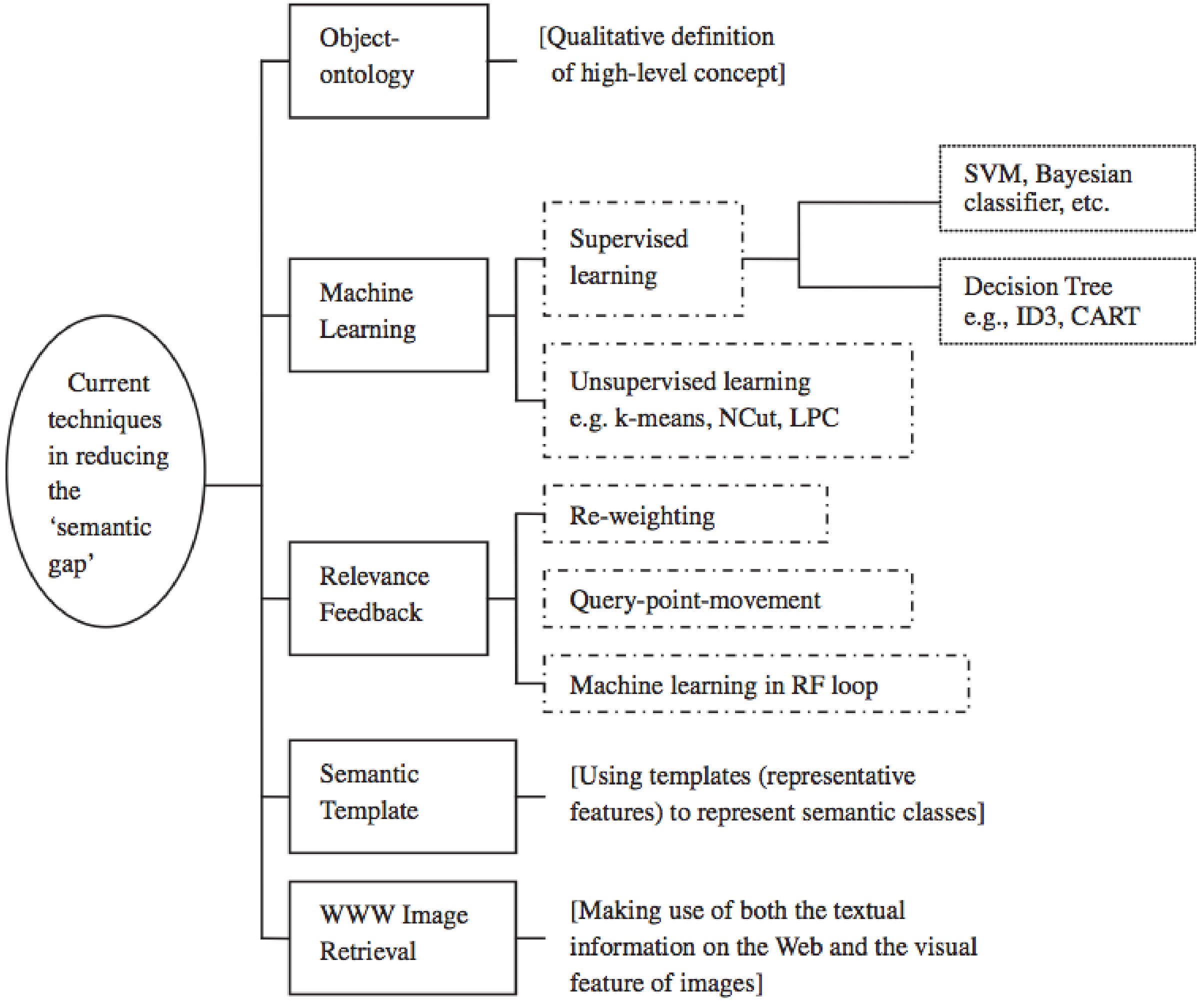
Methodology: Indexing and retrieval

- Most current systems perform retrieval at level 2
  - Low-level image feature extraction (global/regions) → segmentation+characterization
  - Similarity measure
    - Distances between regions → $d(X,Y) = \left(\sum_{i=1}^{p} |x_i - y_i|^r \right)^{1/r}$
    - Distance at image level
      - One-one match: each region in the query image is only allowed to match one region in the target image
      - Many-many match: each region in the query image is allowed to match more than one region in the target image
    - Semantic gap reduction

Methodology: Indexing and retrieval

- Narrowing down the “semantic gap” techniques
  - Object ontology to define high-level concepts
  - Machine learning to associate low-level features with query concepts
  - Relevance feedback to learn users’ intention
  - Generating semantic template to support high-level image retrieval

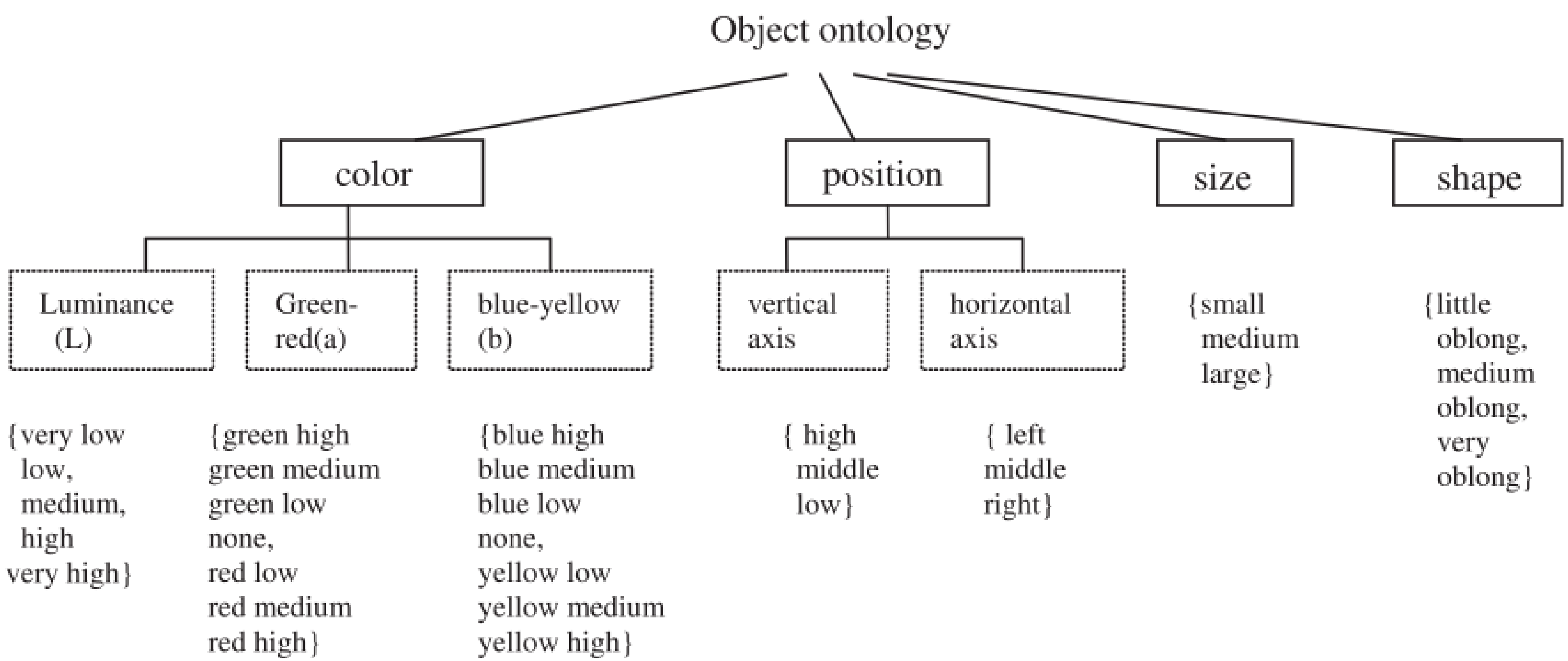
Methodology: Indexing and retrieval



Source: Y. Liu, D. Zhang, G. Lu, and W.-Y. Ma, "A survey of content-based image retrieval with high-level semantics," *Pattern Recognition*, vol. 40, no. 1, pp. 262–282, Jan. 2007.

Methodology: Indexing and retrieval

- Object ontology to define high-level concepts



Source: Y. Liu, D. Zhang, G. Lu, and W.-Y. Ma, “A survey of content-based image retrieval with high-level semantics,” *Pattern Recognition*, vol. 40, no. 1, pp. 262–282, Jan. 2007.

Methodology: Indexing and retrieval

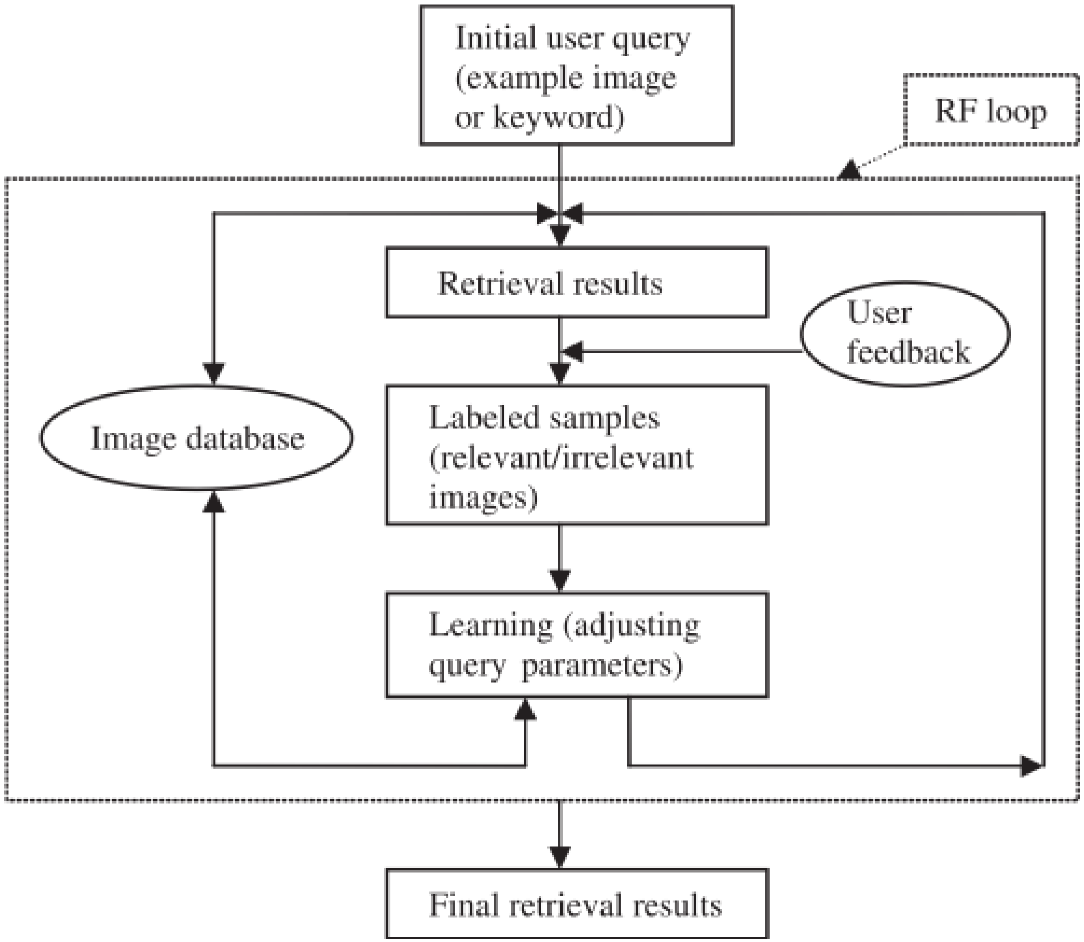
- Machine learning to associate low-level features with query concepts
  - Supervised learning
  - Unsupervised learning
  - Object recognition techniques

Methodology: Indexing and retrieval

- Relevance feedback to learn users'intention
  - Typical scenario
    1. The system provides initial retrieval results through query-by-example, sketch, etc.
    2. User judges the above results as to whether and to what degree, they are relevant (positive examples)/irrelevant (negative examples) to the query.
    3. Machine learning algorithm is applied to learn the user' feedback. Then go back to (2).

Methodology: Indexing and retrieval

Relevance feedback to learn users' intention



Source: Y. Liu, D. Zhang, G. Lu, and W.-Y. Ma, "A survey of content-based image retrieval with high-level semantics," *Pattern Recognition*, vol. 40, no. 1, pp. 262–282, Jan. 2007.

Methodology: Indexing and retrieval

- Generating semantic template (ST) to support high-level image retrieval
  - ST is a map between high-level concept and low-level visual features
  - Different levels of user interaction

TEXT PROCESSING

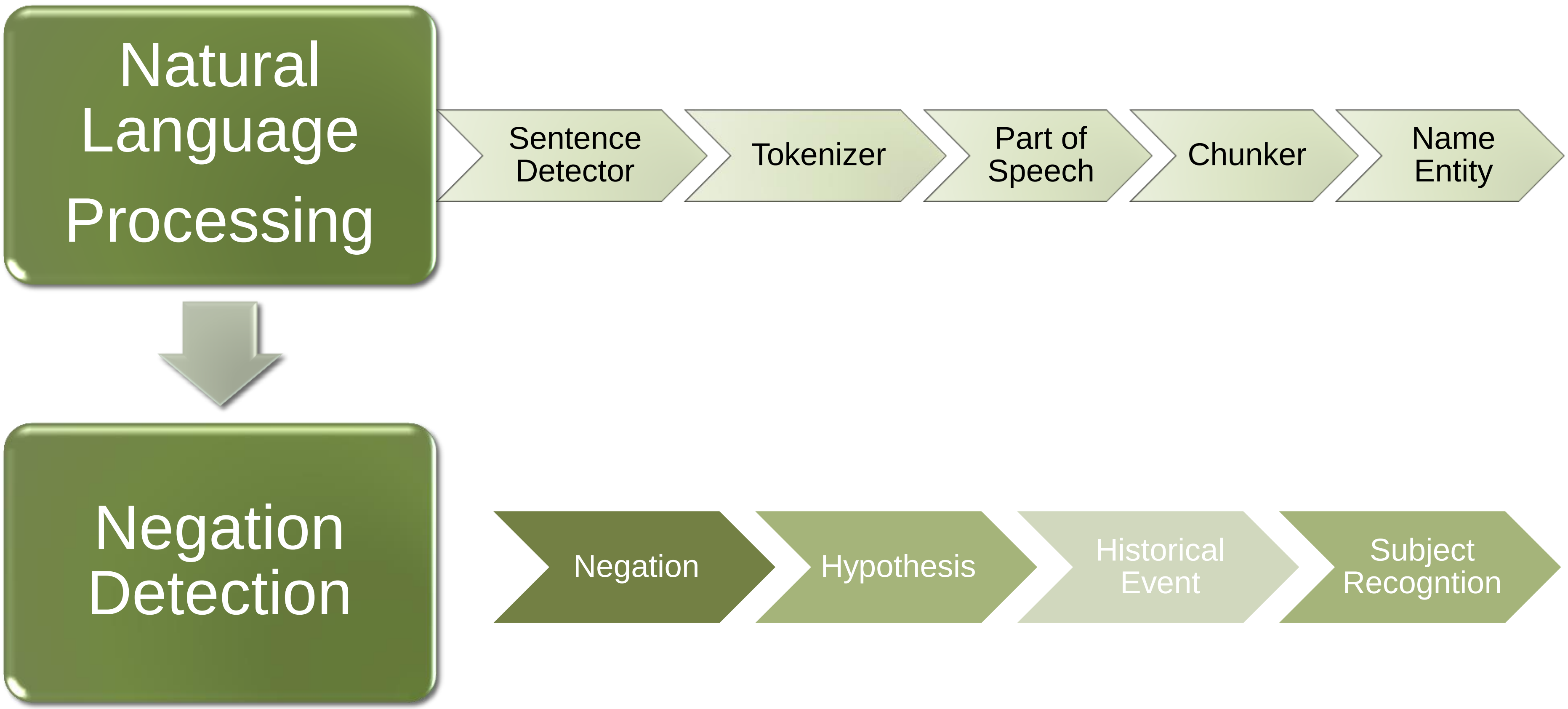
Clinical notes and reports

- Clinical notes contain rich and diverse source of information
- Clinical documents are a valuable source of information for detection and characterization of outbreaks, decision support, recruiting patients for clinical trials, and translational research.
- They contain information regarding signs, symptoms, treatments, and outcomes
- 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:
  - Summary
  - Antecedents (relatives and therapeutical)
  - Tests.
  - judgement
  - treatment

NLP applied to EHR

- Analysis of free text input from clinical reports and patient's history would improve healthcare.
- There are several English-centric tools working towards that goal:

- ✓ Mayo's cTAKES
- ✓ MetaMap
- ✓ MedLee
- ✓ HiTex
- ✓ SNOMED-CT
- ✓ UMLS
- ✓ LOINC


Natural Language Processing
Sentence Detector
Tokenizer
Part of Speech
Chunker
Name Entity
Negation Detection
Negation
Hypothesis
Historical Event
Subject Recogntion

NLP

Paciente varón de 62 años con diagnóstico de carcinoma estadio IV

Paciente varón de 62 años con diagnóstico de carcinoma estadio IV

Paciente varón de 62 años con diagnóstico de carcinoma estadio IV

NN JJ IN NN NNS IN NN IN NN NN NN

Paciente varón de 62 años con diagnóstico de carcinoma estadio IV

NP PP NP PP NP PP NP

NLP training

- Annotated Corpus
- OpenNLP requires to set the values for:
  1. number of iterations: number of times the training procedure should iterate when to find the best the model's parameters;
  2. cut-off: number of times a feature must have been seen in order to be considered into the model.
- Training models:
  - The validation of all the models is done on the basis of a 10-fold cross- validation with 80/20 split
  - precision, recall, accuracy, and F-Measure for trained models

Negation

- Patient's medical records contain valuable clinical information.
- An important feature of the clinical narrative text is that it commonly encloses negation concepts.
- According to Chapman et al. [1], around half of all clinical conditions in narrative reports are negated.

NegEx

- Triggers:
  - definiteExistence,
  - definiteNegatedExistence,
  - historical
- Scope
- Direction

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% specificit 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.

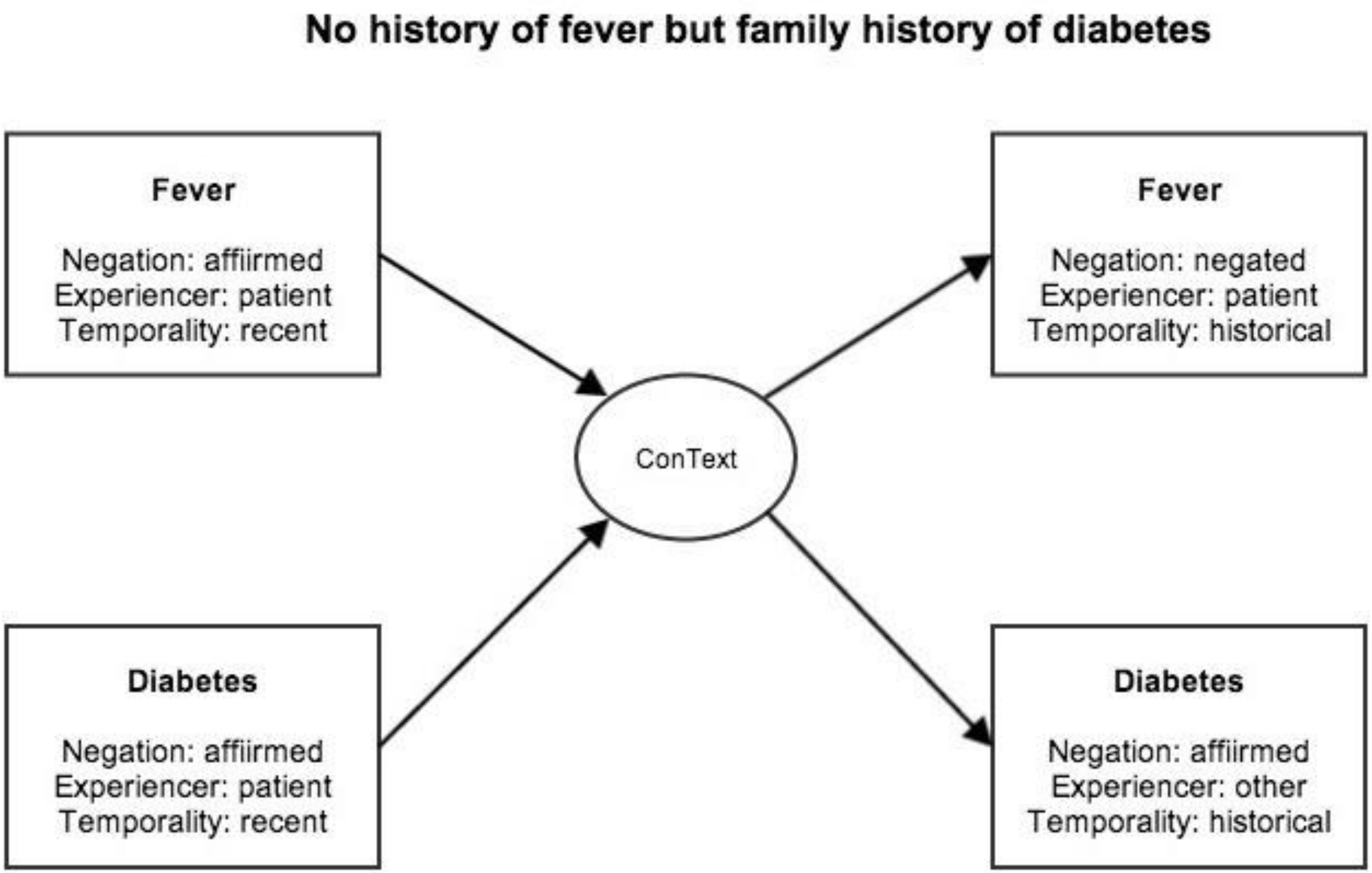
4. Huang Y, Lowe HJ. A novel hybrid approach to automated negation detection in clinical radiology reports. JAMIA 2007

ConText

- **ConText** [1] an extension of the NegEx that uses regular expressions to detect the negation
- determining whether clinical conditions mentioned in clinical reports are:
  - negated: ruled out pneumonia
  - Hypothetical: Patient should return if she develops fever
  - Temporality: historical or recent past history of pneumonia
  - Contextual: experienced by someone other than the patient: family history of pneumonia
- potential to substantially improve precision for information retrieval and extraction from clinical records.
- query for patients with a diagnosis of pneumonia may return false positive records for which pneumoniais mentioned but is negated experienced by a family member or occurred in the past .

[1] Henk Harkema, John N. Dowling, Tyler Thornblade, Wendy W. Chapman, ConText: An algorithm for determining negation,experiencer, and temporal status from clinical reports, Journal of Biomedical Informatics, Volume 42, Issue 5, October 2009, Pages
839-851, ISSN 1532-0464,

ConText


No history of fever but family history of diabetes
Fever
Negation: affiirmed
Experiencer: patient
Temporality: recent
Diabetes
Negation: affiirmed
Experiencer: patient
Temporality: recent
ConText
Fever
Negation: negated
Experiencer: patient
Temporality: historical
Diabetes
Negation: affiirmed
Experiencer: other
Temporality: historical

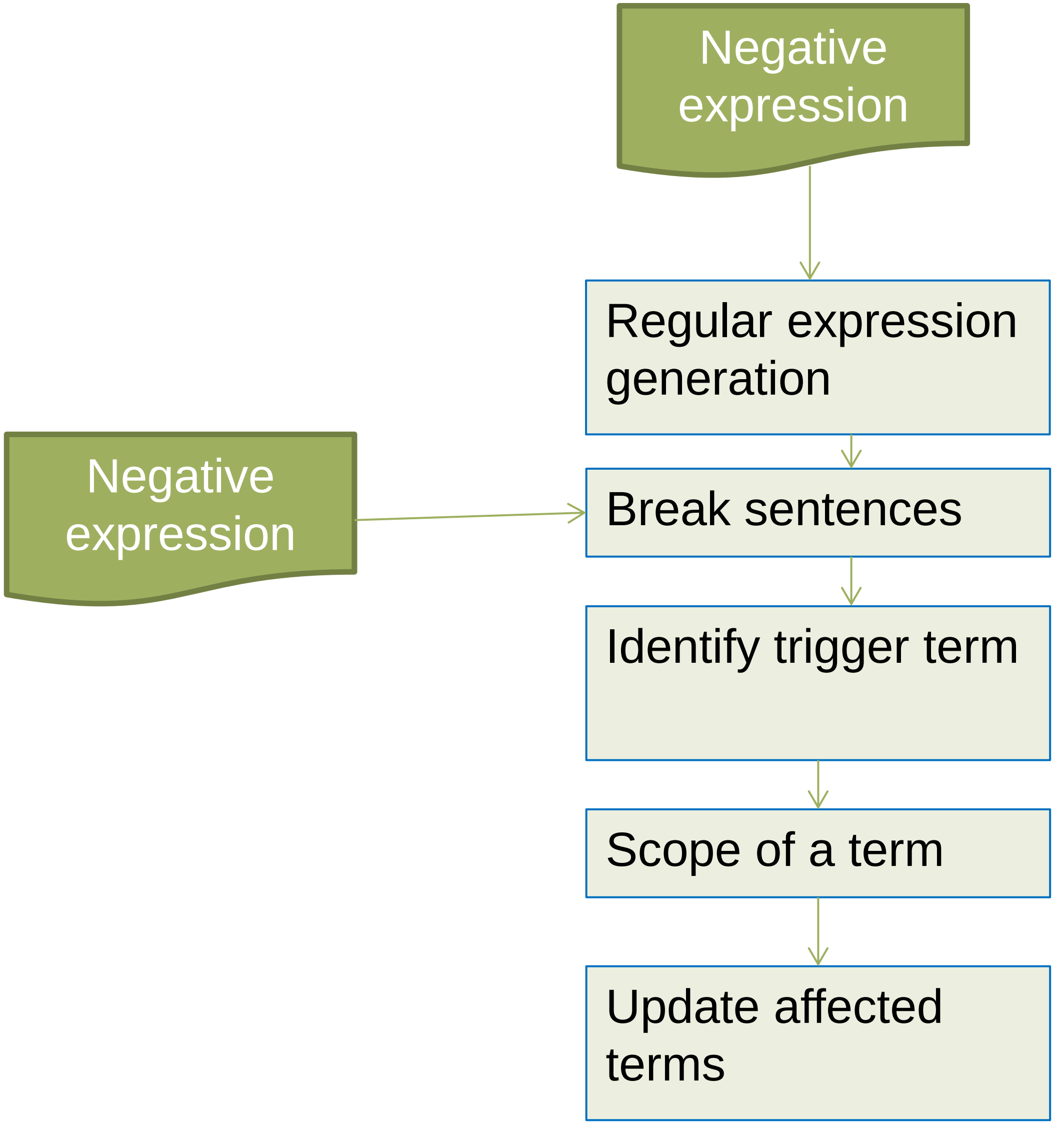
ConText: Generating regular expressions

ConText is based on two types of terms triggers:

- The terms: terms indicating the clinical concept status:
  - denied or affirmed,
  - recent or historical
  - experienced by the patient or otherwise
- in the scope of the term trigger.
- pseudo-trigger terms: Triggers terms resemble but do not work as such terms

- two types of terms depending on their position regarding the concept analyzed terms:
  - Preconcept triggers and triggers postconcept terms.

ContEx



Extends NegEx:

- uses regular expressions to identify the scope of trigger terms that are indicative of negation such as "no" and "ruled out." Any clinical conditions within the scope of a trigger term are marked as negated
- employs a different definition for the scope of trigger terms
- ConText identifies three contextual values in addition to NegEx's negation: hypothetical, historical, and experiencer

ConText: Generating regular expressions

ConText is based on two types of terms triggers:

- The terms: terms indicating the clinical concept status:
  - denied or affirmed,
  - recent or historical
  - experienced by the patient or otherwise
- in the scope of the term trigger.
- pseudo-trigger terms: Triggers terms resemble but do not work as such terms

- two types of terms depending on their position regarding the concept analyzed terms:
  - Preconcept triggers and triggers postconcept terms.

ConText: triggers

- Identify all trigger terms:
  - “no” and “denies,”
  - for hypothetical, “if” and “should,”
  - for historical, “history” and “status post,”
  - and for other, “family history” and “mother’s.”
  - The total number of trigger terms used by the current version of ConText is: 143 for negated, 10 for historical, 11 for hypothetical, and 26 for other

ConText: pseudo-triggers

- pseudo-triggers
  - terms that contain trigger terms but do not act as contextual property triggers
  - To avoid false positives, “History exam” is includedin the list of pseudo-triggers for historical.
  - In the current version of ConText there are 17 pseudo-triggers for negated (e.g., “no increase,” “not cause”), 17 pseudo-triggers for historical (e.g., “social history,” “poor history”), four pseudo-triggers for hypothetical (e.g., “if negative,” “know if”), and 18 pseudo-triggers for other (e.g., “by her husband,” “by his brother”)

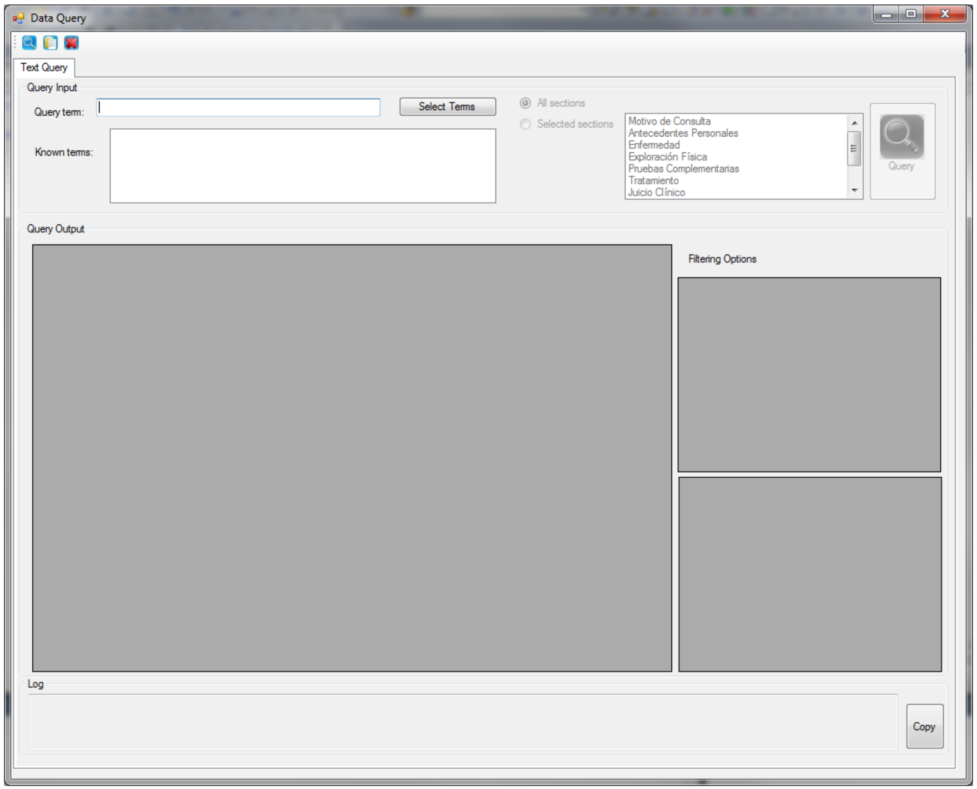
Algorithm

- Mark up all trigger terms, pseudo-trigger terms, and termination terms in the sentence.
- Iterate through the trigger terms in the sentence from left to right:
  - If the trigger term is a pseudo-trigger term, skip to the next trigger term.
  - Otherwise, determine the scope of the trigger term and assign the appropriate contextual property value to all indexed clinical conditions within the scope of the trigger term.

ConText: triggers

- Identify all trigger terms:
  - “no” and “denies,”
  - for hypothetical, “if” and “should,”
  - for historical, “history” and “status post,”
  - and for other, “family history” and “mother’s.”
  - The total number of trigger terms used by the current version of ConText is: 143 for negated, 10 for historical, 11 for hypothetical, and 26 for other
- pseudo-triggers
  - terms that contain trigger terms but do not act as contextual property triggers
  - To avoid false positives, “History exam” is included in the list of pseudo-triggers for historical.
  - In the current version of ConText there are 17 pseudo-triggers for negated (e.g., “no increase,” “not cause”), 17 pseudo-triggers for historical (e.g., “social history,” “poor history”), four pseudo-triggers for hypothetical (e.g., “if negative,” “know if”), and 18 pseudo-triggers for other (e.g., “by her husband,” “by his brother”)

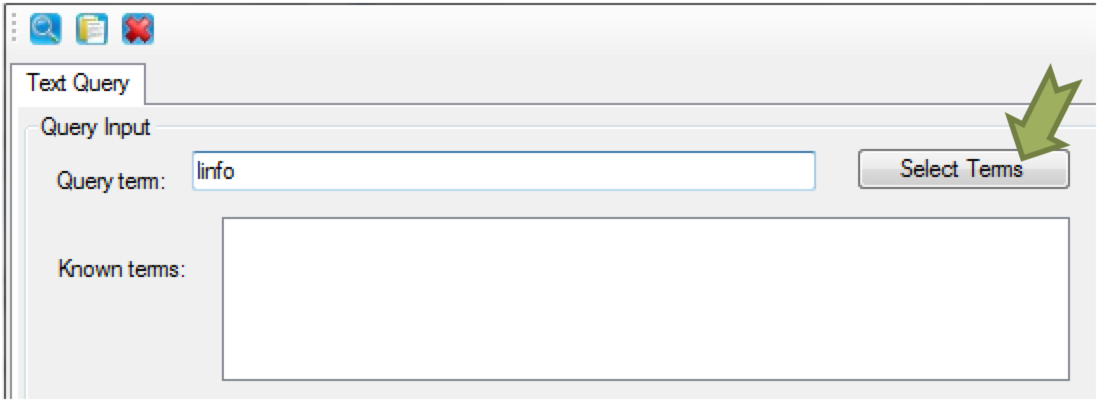
Querying indexed data

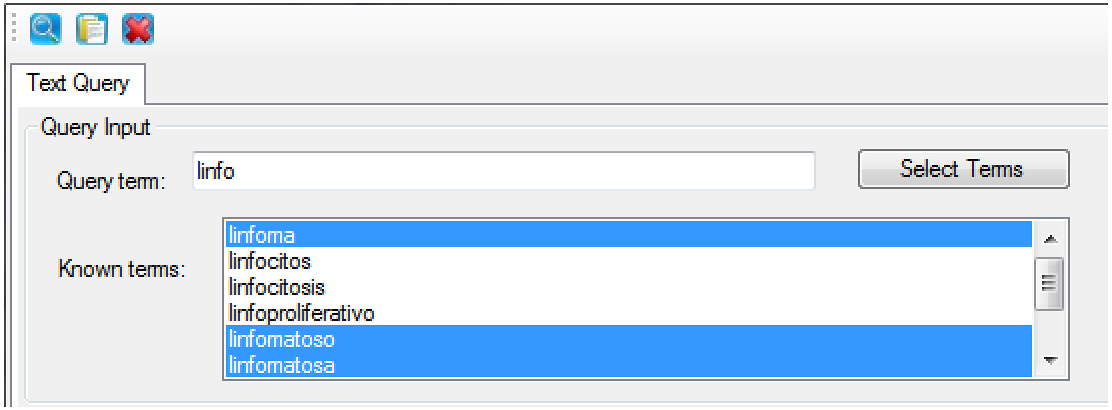


Basic interface functionality filtering search results and reporting of history.

Limitations:
Partial data available.
No definition of terms apply.
Limitations syntactic and semantic indexing terms.

Insert query term



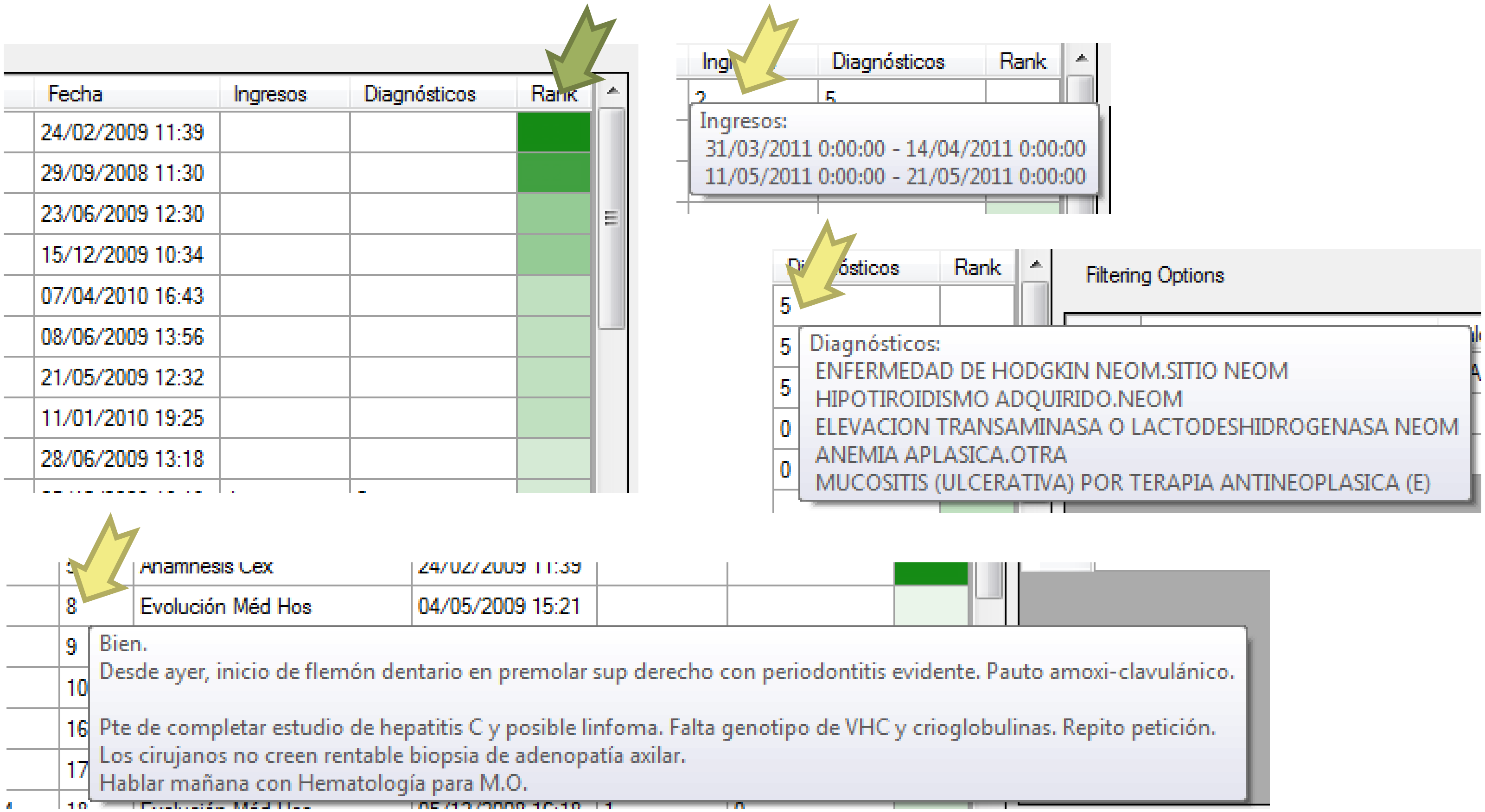


Simple results

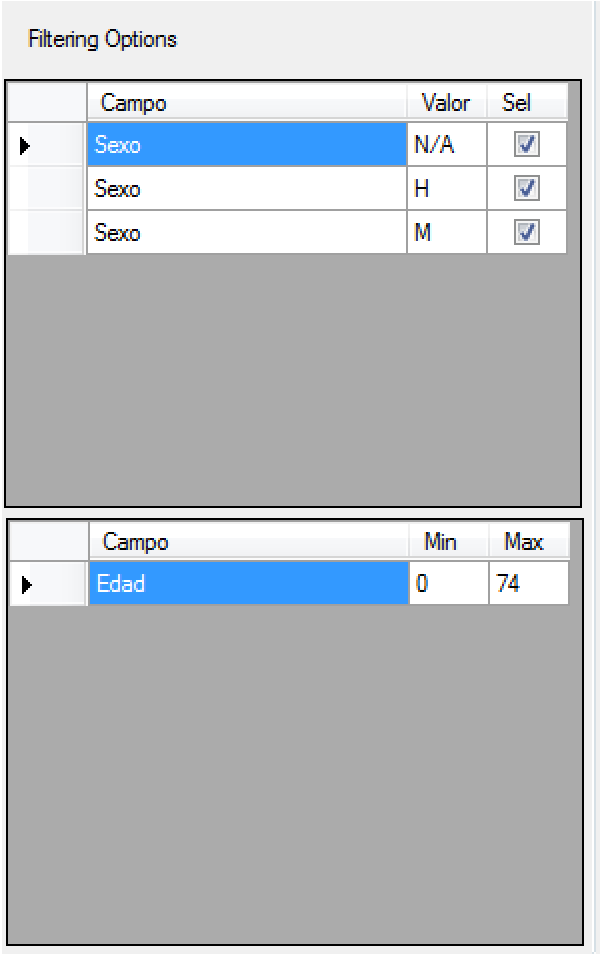
Query Output

Historial	Sexo	Edad	Nota	Tipo de Nota	Fecha	Ingresos	Diagnósticos	Rank
			20	ITC Hematología	21/05/2009 12:32			
			38	Antecedentes Personales	15/12/2009 10:27			
			43	Evolución Méd Hos	18/12/2009 14:08			
			34	ITC Hematología	17/11/2009 14:34			
			8	Evolución Méd Hos	04/05/2009 15:21			
38766222	H	47	84	Evolución Méd Cex	30/09/2009 7:07	2	5	
			78	Pr Complement Cex	24/09/2009 10:18			
			68	Anamnesis Cex	19/12/2009 1:00			
			72	Comentario Enf Hos	11/01/2010 19:25			
			97	Anamnesis Cex	17/03/2009 12:31			
			39	Comentario Enf Hos	14/12/2009 21:04			
			29	Evolución Méd Hos	19/11/2009 14:00			
			16	Pr Complement Hos	18/12/2009 13:53			
			95	Res situación Cex	21/10/2008 15:16			
			69	Anamnesis Hos	25/06/2009 13:32			
			66	Anamnesis Cex	18/06/2009 9:32			
			79	Tratamiento Cex	02/09/2009 10:19			
			23	Comentario Enf Hos	15/11/2009 7:20			
			5	Anamnesis Cex	24/02/2009 11:39			
			96	Nota de Urgencias	12/12/2008 23:56			

Make information easy accesible



Filtering



Interactive filtering data:
Initially on history data.
It is possible to extend it to the income data and diagnostics.

You can include aggregate information on selected items / removed, to help filter run:
Diagnostic statistics.
Visual presentation of histograms.

Tags cloud


aprendizaj complet document cni anual
are conclusion domin frontal distribu clinic anticipacion
asoci consej dra frontoparietal localizacion form disemin cerez ambas
autosom control ecograf general malform otra levement fla dificultad cerebel alteracion
corial entid genet manch paramagnet realizacion organ lev firm diagnostic carm alta
epilepsi germinal manej penetr recomiend sid rar oftalmolog inter fetal diagnost carcinom
cortical manifest perd recurrent siguient bilateral sex quist ofrec innecesari fenomen
medi personal refier simetr cas heterocigosis antecedent sensibil puls ocurr
subependimari centr hiperecogen progres habitual analisis semioval observ
comentari identific rutinari derech progenitor gonzalez actual
seguimient enfermed cerebral dat predomini exon visual
imag hered angiomiolipom blanc person
hiperintens esclerosis secuenci
relacion imagen sol vist semestral pulmonar
sufr terci evalu nodul renal portador vez
teres indic trastorn fech
vascular seman progrnitor numer
hamartom realiz tuber lesion riesg
izquierd tres herman
variabl secundari problem nov
padr sustanci mutacion compat
comport lineal tsc
tsc2 prueb estudi individu inform alter vari
nombr
piel torax localiz variant solicit gen tambi pacient ecogen proband
transmit contr migracion varon son corticosubcortical normal
piern remit tratamient deb moment afect neuronal displas validacion ningun
plan requier troncoencefal detect nacimient discret urolog secuenciacion ingres
escuel gestacion men pol rest tubers diaz unid rutin probabl neurolog inferior alons
cost espald grand miguel posibil result tumor retinian present negat inclu familiar aisl
craneal especial hallazg mosaic posibl resum presenci necesari histori famili desconoc
basal depend estas hamartomat mostr prenatal multipl hipointens expres descendient
acron benefici dermatolog este hemisferi muj hij exam descendent caracteriz adolescent
administracion biopsi desarroll estructur hereditari evaluacion descart captacion

Conclusions

- EHR analysis and evidence-based decisions in hospitals need the adoption of this technologies.
- Efforts in adopting NLP techniques in Biomedicine should be done.
- Image annotation techniques are required
- Integration of image annotation and text processing

POLITÉCNICA
"Ingeniamos el futuro"
CAMPUS DE EXCELENCIA INTERNACIONAL

Conclusions

CTB centro de tecnología biomédica

- Improvement of NLP process
- Improvement of negation detection algorithms to include more contextual information.
- Generation of new algorithms applied to clinical conditions and their relationships.
- Application of data mining techniques to extract knowledge from the system.

Conclusions

- Health domain is generating huge of complex data
- Integrated methods (hw and sw) are required
- Mining clinical notes and Automatic Image annotation (AIA) very challenging research area. There are several major issues:

    1.- High dimensional feature analysis.

    2.- How to build an effective annotation model?

    3.- How to rank images/texts within each of the categories?

    4.- Lack of standard vocabulary and taxonomy for annotation.

ACKNOWLEDGEMENTS

Research Projects

- PROJECTS
  - Rethink Big
  - Resilience 2050
  - Estudio de re-análisis de imágenes y correlación entre los cambios metabólicos objetivados por PET/CT y las mutaciones conocidas en el cáncer de pulmón no célula pequeña (CPNCP)
- COOPERATION
  - StreaMED "Data Mining and Stream Mining for Epidemiological Studies on the Human Brain“ with Otto VonVericke.Magdemburg
- Hospitals
  - Hospital Puerta de Hierro
  - Hospital de la Princesa

People

- MIDAS research group
  - Consuelo Gonzalo
  - Jose María Peña
  - Roberto Costumero
  - Angel Mario Garcia
- Cooperation
  - Myra Spiliopolou. Magdemburg
  - Fernando Maestu
- Hospitals:
  - Jorge Gomez Zamora
  - Juan Luis Cruz Bermudez
  - Mariano Provencio

THANKS

LITERATURE

Text processing

Henk Harkema, John N. Dowling, Tyler Thornblade, Wendy W. Chapman, ConText: An algorithm for determining negation,experiencer, and temporal status from clinical reports, Journal of Biomedical Informatics, Volume 42, Issue 5, October 2009, Pages

839-851, ISSN 1532-0464

Aronow DB, Fangfang F, Croft WB. Ad hoc classification of radiology reports. JAMIA 1999:393-411

Chapman et al. A simple algorithm for identifying negated findings and diseases in discharge summaries. JBI 2001:301-10.

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.

Huang Y, Lowe HJ. A novel hybrid approach to automated negation detection in clinical radiology reports. JAMIA 2007

Chapman W., Bridewell W., Hanbury P., Cooper G.F., Buchanan B. A Simple Algorithm for identifying negated findings and diseases in discharge summaries. Journal of Biomedical Informatics 34, 301-310 (2001)

Henk Harkema, John N. Dowling, Tyler Thornblade, Wendy W. Chapman. ConText: An algorithm for determining negation, experiencer, and temporal status from clinical reports. Journal of Biomedical Informatics 42 (2009) 839–851

Morante R., Liekens A., Daelemans W. Learning the Scope of Negation in Biomedical Texts. Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pages 715–724, Honolulu, October 2008. © 2008 Association for Computational Linguistics

Skeppstedt Maria. Negation detection in Swedish clinical text: An adaption of NegEx to Swedish. Journal of Biomedical Semantics 2011, 2(Suppl 3):S3Avalilable at http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3194175/pdf/2041-1480-2-S3-S3.pdf

Meystre S.M., Savova G.K., Kipper-Schuler K.C., Hurdle J.F. Extracting Information from Textual Documents in Electronic Health Record: A Review of Recent Research. IMIA Yearbook of Medical Informatics 2008, 138-154

Image Processing

- D. Zhang, M. M. Islam, and G. Lu, “A review on automatic image annotation techniques,” Pattern Recognition, vol. 45, no. 1, pp. 346–362, Jan. 2012.+
- Y. Liu, D. Zhang, G. Lu, and W.-Y. Ma, “A survey of content-based image retrieval with high-level semantics,” *Pattern Recognition*, vol. 40, no. 1, pp. 262–282, Jan. 2007.
- Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Su x0308 sstrunk, S.: SLICSuperpixels Compared to State-of-the-Art Superpixel Methods. Pattern Analysis and Machine Intelligence, IEEE Transactions on 34(11) (2012) 2274{22822. Hay, G.J., Castilla, G.: Object-based image analysis: strengths, weaknesses, opportunities and threats (SWOT). International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences 36(4) (2006)
- Ren, X., Malik, J.: Learning a classication model for segmentation. In: Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on, IEEE (2003)10{17 vol.1
- Sluimer, I., Schilham, A., Prokop, M., van Ginneken, B.: Computer analysis ofcomputed tomography scans of the lung: a survey. IEEE Transactions on MedicalImaging 25(4) (2006) 385{405
- Zhang, L., Homan, E., Reinhardt, J.: Atlas-driven lung lobe segmentation in volumetric x-ray ct images. Medical Imaging, IEEE Transactions on 25(1) (Jan 2006) 1{16
- Moltz, J., Bornemann, L., Kuhnigk, J.M., Dicken, V., Peitgen, E., Meier, S., Bolte,H., Fabel, M., Bauknecht, H.C., Hittinger, M., Kiessling, A., Pusken, M., Peitgen,H.O.: Advanced segmentation techniques for lung nodules, liver metastases, and enlarged lymph nodes in ct scans. Selected Topics in Signal Processing, IEEE Journal of 3(1) (Feb 2009) 122{134
- Hardie, R., Rogers, S., Wilson, T., Rogers, A.: Performance analysis of a new computer aided detection system for identifying lung nodules on chest radiographs. Medical Image Analysis 12(3) (2008) 240|-258
- Nakagomi, K., Shimizu, A., Kobatake, H., Yakami, M., Fujimoto, K., Togashi, K.:Multi-shape graph cuts with neighbor prior constraints and its application to lung segmentation from a chest CT volume . Medical Image Analysis 17(1) (2013) 62{ 77

Image Processing

- Heimann, T., van Ginneken, B., Styner, M.A., Arzhaeva, Y., Aurich, V., Bauer,C., Beck, A., Becker, C., Beichel, R., Bekes, G., Bello, F., Binnig, G., Bischof, H., Bornik, A., Cashman, P., Chi, Y., Cordova, A., Dawant, B.M., Fidrich, M.,Furst, J.D., Furukawa, D., Grenacher, L., Hornegger, J., Kainmuller, D., Kitney, R.I., Kobatake, H., Lamecker, H., Lange, T., Lee, J., Lennon, B., Li, R., Li, S.,Meinzer, H.P., Nemeth, G., Raicu, D.S., Rau, A.M., van Rikxoort, E.M., Rousson, M., Rusko, L., Saddi, K.A., Schmidt, G., Seghers, D., Shimizu, A., Slagmolen,P., Sorantin, E., Soza, G., Susomboon, R., Waite, J.M., Wimmer, A., Wolf, I.: Comparison and Evaluation of Methods for Liver Segmentation From CT Datasets.IEEE Transactions on Medical Imaging 28(8) (2009) 1251{1265
- Zhang, X., Tian, J., Deng, K., Wu, Y., Li, X.: Automatic liver segmentation using a statistical shape model with optimal surface detection. Biomedical Engineering, IEEE Transactions on 57(10) (2010) 2622{2626
- D Kainmueller, T.L., H.Lamecker: Shape Constrained Automatic Segmentation of the Liver based on a Heuristic Intensity Model. In: 3D Segmentation in the Clinic:
- A Grand Challenge (MICCAI), 2007 Workshop on. (2007)
- T. Okada, R. Shimada, Y.S.M.H.K.Y.M.N.Y.C.H.N., Tamura, S.: Automated segmentation of the liver from 3D CT images using probabilistic atlas and multilevel statistical shape model. In: 3D Segmentation in the Clinic: A Grand Challenge (MICCAI), 2007 Workshop on. (2007)
- Heimann, T., Meinzer, H.P., Wolf, I.: A statistical deformable model for the segmentation of liver CT volumes. 3D Segmentation in the clinic: A grand challenge (2007) 161{166
- Chitiboi, T., Hennemuth, A., Tautz, L., Stolzmann, P., Donati, O.F., Linsen, L., Hahn, H.K.: Automatic detection of myocardial perfusion defects using objectbased myocardium segmentation. In: Computing in Cardiology Conference (CinC), 2013. (2013) 639{642
- Schwier, M., Moltz, J.H., Peitgen, H.O.: Object-based analysis of CT images for automatic detection and segmentation of hypodense liver lesions. International Journal of Computer Assisted Radiology and Surgery 6(6) (April 2011) 737{747
- Massoptier, L., Casciaro, S.: A new fully automatic and robust algorithm for fast segmentation of liver tissue and tumors from CT scans. European Radiology 18(8) (March 2008) 1658{1665
- Jiang, H., Tang, F., Zhang, X.: Liver cancer identication based on PSO-SVM model. In: Control Automation Robotics & Vision (ICARCV), 2010 11th International Conference on. (2010) 2519{2523
- Ling, H., Zhou, S.K., Zheng, Y., Georgescu, B., Suehling, M., Comaniciu, D.: Hierarchical, learning-based automatic liver segmentation. In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. (2008) 1{8