

Biomedical Semantics in the Big Data Era

Workshop of IMIA WG 6
'Language and Meaning in Biomedicine' (LaMB)
MEDINFO 2015 - São Paulo, Brazil



Speakers

- ▶ Tomasz Adamusiak (Thomson Reuters, Boston, MA, USA)
- ▶ Ronald Cornet (Academic Medical Center University of Amsterdam, Amsterdam, The Netherlands & Linköping University, Linköping, Sweden)
- ▶ Jianying Hu (IBM T. J. Watson Research Center, Yorktown Heights, NY, USA)
- ▶ Stephane Meystre (University of Utah, Salt Lake City, Utah, USA)
- ▶ Patrick Ruch (University of Applied Sciences Western Switzerland, Geneva, Switzerland)
- ▶ Stefan Schulz (Medical University of Graz, Graz, Austria)

Comment Stefan: I removed academic degree because of difficult comparability across countries

Agenda

- ▶ Introduction - Ronald Cornet
- 1. From free text to ontology - Stephane Meystre
- 2. Bridging natural and formal languages for representing knowledge and information - Stefan Schulz
- 3. Deep question-answering for biomedical decision support - Patrick Ruch
- 4. Feature extraction for predictive modeling - Jianying Hu
- 5. Semantic technology for knowledge discovery - Tomasz Adamusiak
- ▶ Overall discussion - all
- ▶ Round-up

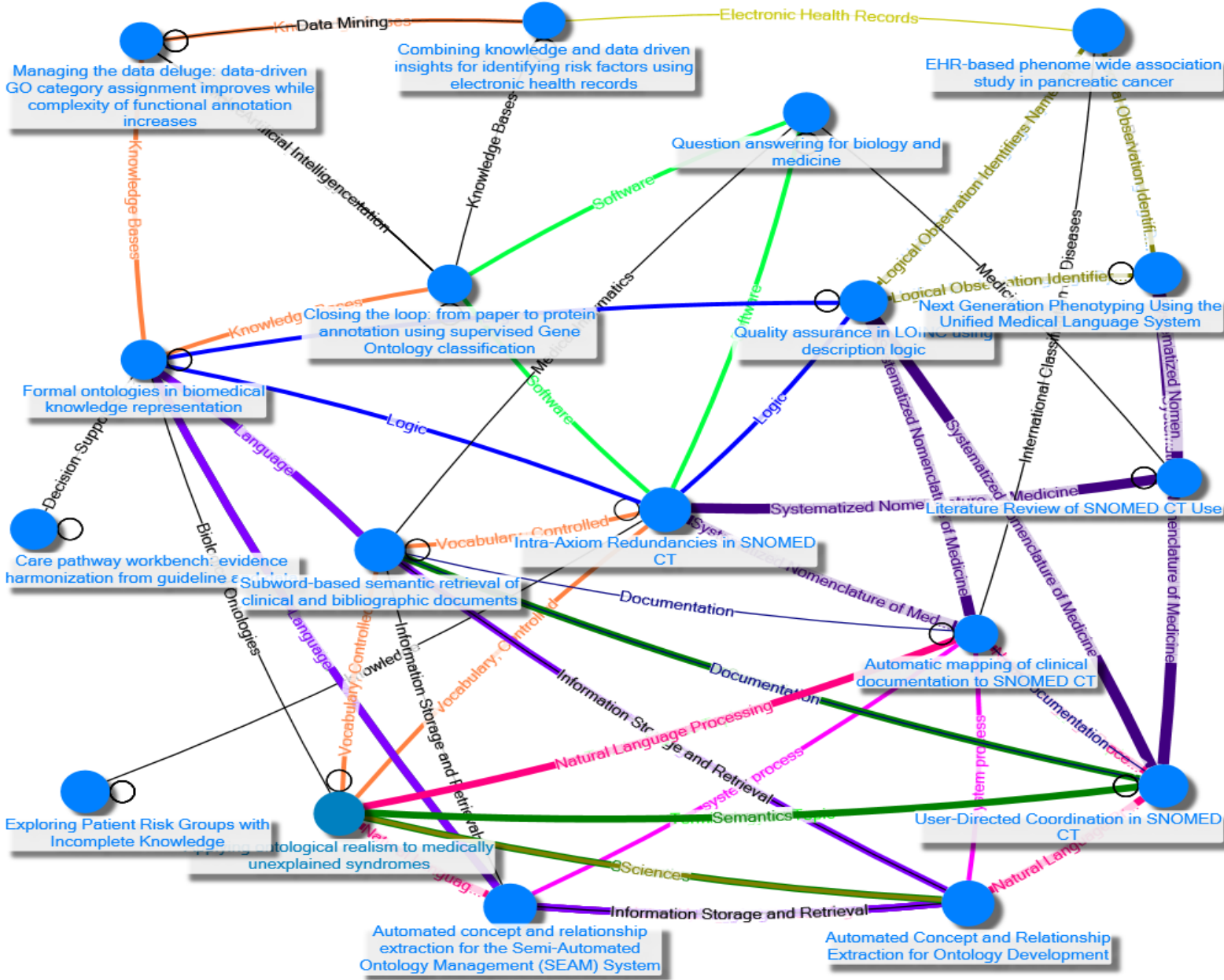
Topics to be discussed

- Applying ontological realism to medically unexplained syndromes
- Automated Concept and Relationship Extraction for Ontology Development
- Automated concept and relationship extraction for the Semi-Automated Ontology Management (SEAM) System
- Automatic mapping of clinical documentation to SNOMED CT
- Care pathway workbench: evidence harmonization from guideline and data
- Closing the loop: from paper to protein annotation using supervised Gene Ontology classification
- Combining knowledge and data driven insights for identifying risk factors using electronic health records
- EHR-based phenome wide association study in pancreatic cancer
- Exploring Patient Risk Groups with Incomplete Knowledge
- Formal ontologies in biomedical knowledge representation
- Intra-Axiom Redundancies in SNOMED CT
- Literature Review of SNOMED CT Use
- Managing the data deluge: data-driven GO category assignment improves while complexity of functional annotation increases
- Next Generation Phenotyping Using the Unified Medical Language System
- Quality assurance in LOINC using description logic
- Question answering for biology and medicine
- Subword-based semantic retrieval of clinical and bibliographic documents
- User-Directed Coordination in SNOMED CT

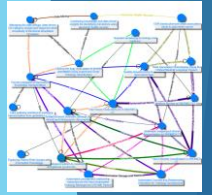
Key Concepts

<ul style="list-style-type: none">• Algorithms• Artificial Intelligence	<ul style="list-style-type: none">• Linguistics• Logic
<ul style="list-style-type: none">• Automatic Data Processing• Bayes• Bibliometrics• Biological Ontologies• Biological Science Disciplines• Biology	<ul style="list-style-type: none">• Logical Observation Identifiers Names and Codes• Meaningful Use• Medical Informatics• Medical Informatics Applications• medical record• Medical Record Linkage
<ul style="list-style-type: none">• Case Management• Classification• Computational Biology• Computer Systems• Computers• Critical Pathways• Current Procedural Terminology• data• Data Collection• Data Mining• Databases, Bibliographic• Databases, Genetic	<ul style="list-style-type: none">• Medical Records Systems, Computerized• Medicine• MEDLINE• Molecular Sequence Annotation• Multilingualism• Natural Language Processing• predictive model• Programming Languages• Publishing• PubMed• Quality Control• Quality Improvement
<ul style="list-style-type: none">• Decision Support Systems, Clinical• Decision Support Techniques• Documentation• Electronic Health Records• Gene Ontology	<ul style="list-style-type: none">• RiSk Group Analysis• ROC Curve• RxNorm• Science• Search Engine
<ul style="list-style-type: none">• Healthcare Common Procedure Coding System• Information Science	<ul style="list-style-type: none">• Semantics• Software
<ul style="list-style-type: none">• Information Storage and Retrieval	<ul style="list-style-type: none">• system process
<ul style="list-style-type: none">• Information Systems• Intelligence	<ul style="list-style-type: none">• Systematized Nomenclature of Medicine• Systems Biology
<ul style="list-style-type: none">• International Classification of Diseases• knowledge• Knowledge Bases• Language	<ul style="list-style-type: none">• Terminology as Topic• Unified Medical Language System• Vocabulary, Controlled• Workflow

Topics and key concepts



First presentation



Biomedical Semantics in the Big Data Era

Stefan Schulz

Bridging natural and formal
languages for representing
knowledge and information



Workshop of IMIA WG 6
'Language and Meaning in Biomedicine' (LaMB)

MEDINFO 2015 - São Paulo, Brazil

The promise of Big Data (?)

- ▶ *"Today companies like Google, which have grown up in an era of massively abundant data, don't have to settle for wrong models. Indeed, they don't have to settle for models at all. (...) Forget taxonomy, ontology, and psychology."*

Chris Anderson *Wired Magazine*, 2008

Translate

The screenshot shows the Google Translate interface. At the top, the word "Translate" is written in red. Below it, there are language selection buttons for "English", "French", "Spanish", and "Detect language". A text input box contains the English text: "She was admitted after being found to be demented and unclear if this was a new or old diagnosis. Basic labs for dementia were sent including TSH, B12, folate which were normal". Below the input box are icons for voice input, speaker output, and keyboard input. To the right of the input box are icons for Google+, a star, and a refresh button. Below the input box, there are language selection buttons for "French", "English", and "Portuguese", followed by a blue "Translate" button. The output box shows the Portuguese translation: "Ela foi internada depois de ter sido encontrado para ser demente e não está claro se este foi um diagnóstico novo ou velho. Laboratórios básicos para demência foram enviados incluindo TSH, B12, ácido fólico, que eram normais". At the bottom of the output box are icons for a star, a copy icon, a speaker icon, and a "Wrong?" link.

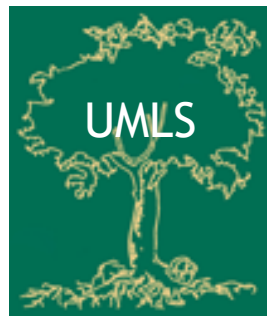
The rise of semantic standards and specifications

Ontologies

Terminologies

Information Models

- IHTSDO delivering
- SNOMED CT[®]
- the global clinical terminology



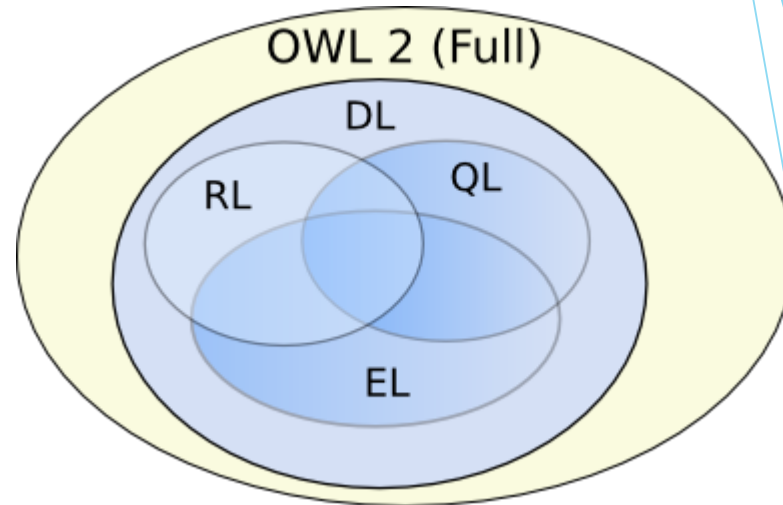
*open***EHR**



OBO



Standardised representation and reasoning formalisms

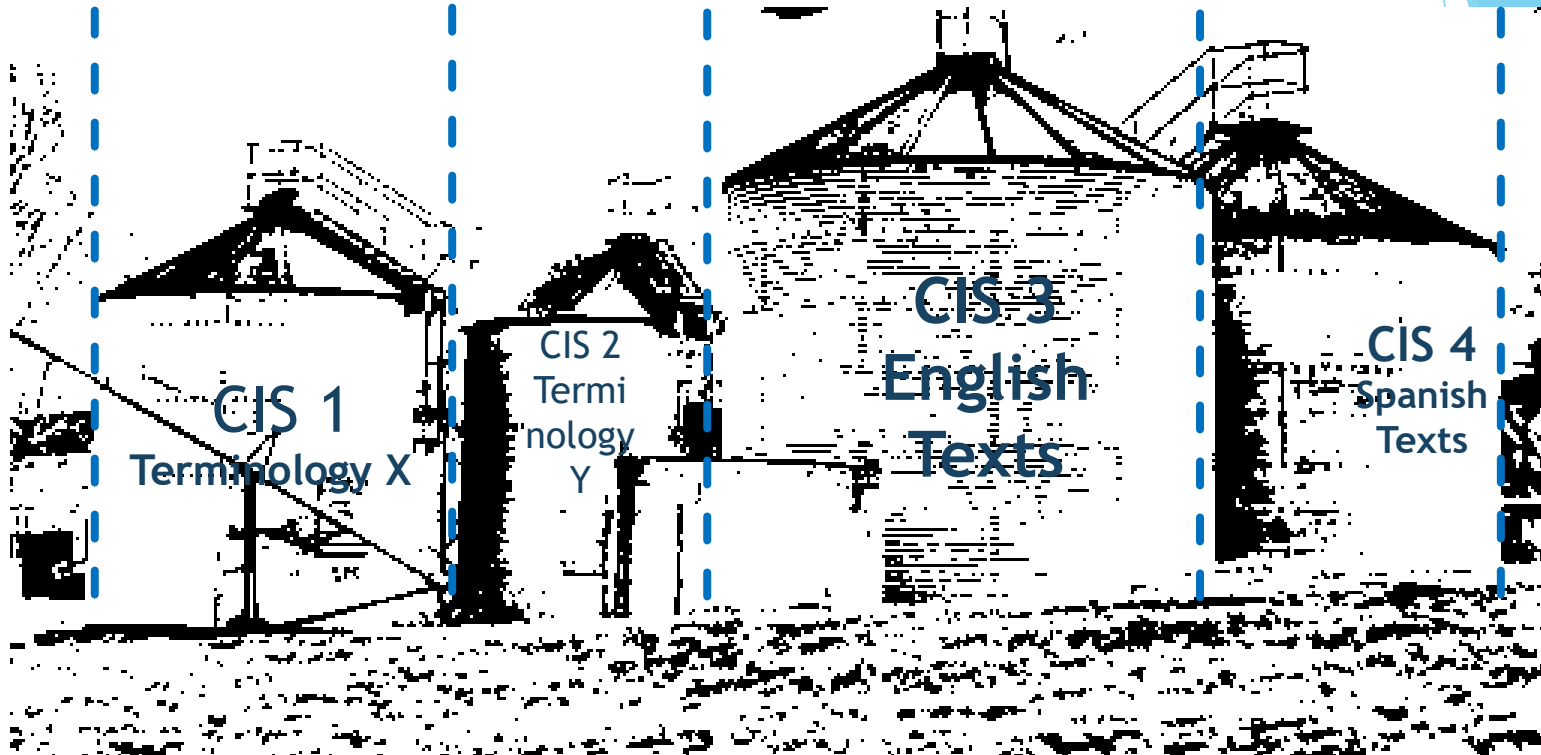


ObservationResult and *isAboutQuality* only (*MassIntake* and *inheresIn* some *CigaretteTobaccoSmokingSituation* and *projectsOnto* some (*ValueRegion* and *isRepresentedBy* only (hasInformationAttribute some *perDay* and hasValue some int[>=10])))
SubClassOf *InformationItem* and *isAboutSituation*
HeavyCigaretteTobacco Situation

Persistence of free text narratives in EHRs

Patient is a 80 y/o female with hx of CAD, DM, HTN, left PICA stroke who presented to the ED after a fall. She was admitted after being found to be demented and unclear if this was a new or old diagnosis. Basic labs for dementia were sent including TSH, B12, folate which were normal. MRI revealed a meningioma and old PICA infarct. Likely diagnosis is Alzheimer's Disease. She was started on donepezil and Quetiapine. PT/OT evaluated her and felt that she was safe to be d/c home with services.

Persistence of data silos: No interoperability



- ▶ Proprietary vocabularies / data dictionaries
- ▶ Proprietary information templates
- ▶ Different natural languages
- ▶ Legacy systems that obviate data exchange

Barriers to Semantic Interoperability

- ▶ Vocabularies, ontologies, information models:
 - ▶ Conflicting and overlapping models of meaning and use
 - ▶ Lack of ontological grounding
 - ▶ Confuse and ambiguous naming
- ▶ Representation and reasoning mechanisms
 - ▶ Difficult to learn and to apply
 - ▶ Performance issues, computational complexity
 - ▶ Lack of industry-standard tools
- ▶ Natural language content
 - ▶ Idiosyncratic language (abbreviated, ungrammatical)
 - ▶ Context dependence

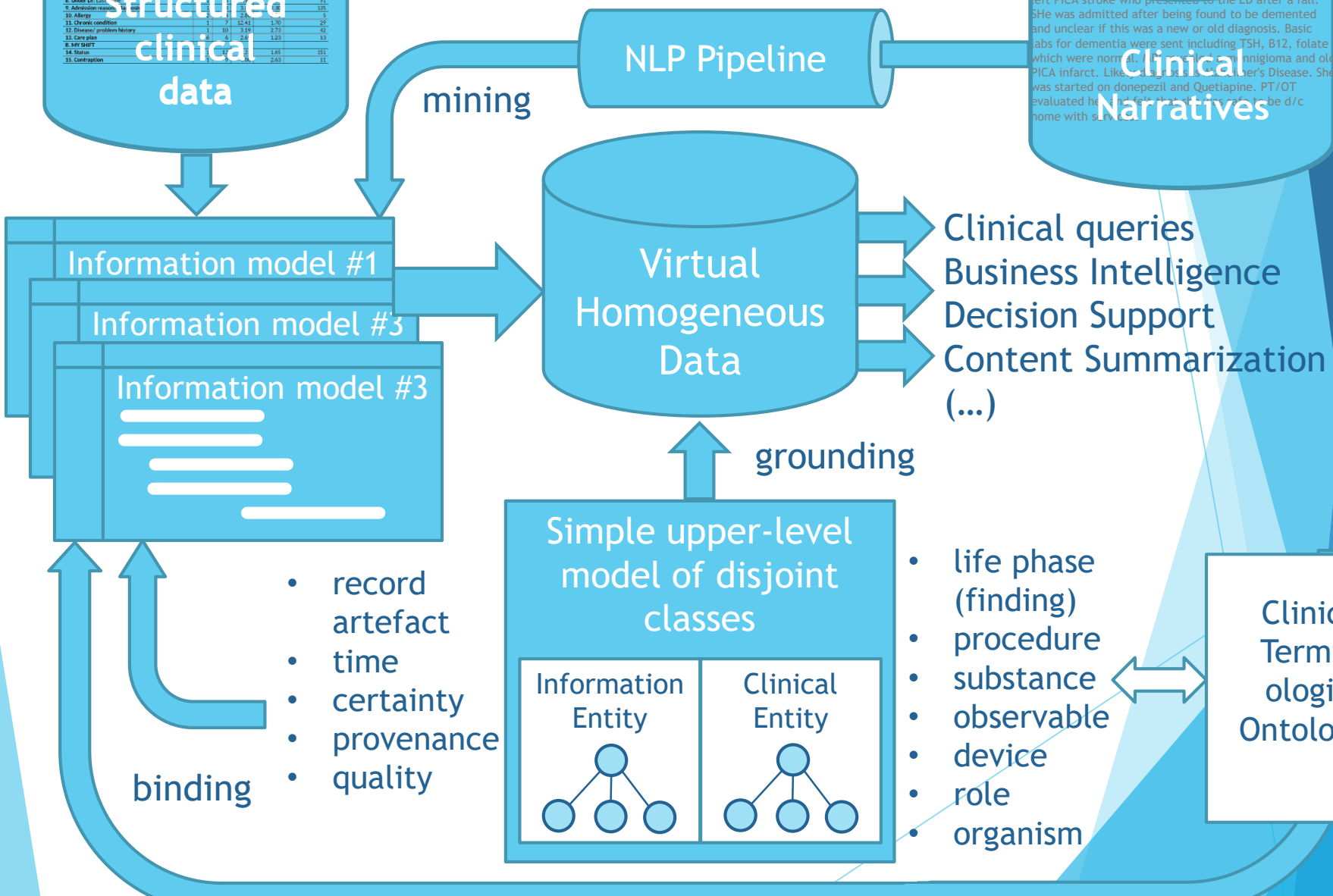
The vision of Semantic Big Data

Structured clinical data

CATEGORY	LENGTH OF A HIGHLIGHTED TEXT SNIPPET (WORDS)				NUMBER OF HIGHLIGHTED TEXT SNIPPETS
	Min	Max	Mean	Standard deviation	
...PATIENT INTRODUCTION					
1. Given name/ initials	1	2	1.12	0.33	107
2. Last name	1	2	1.01	0.29	99
3. Age in years	1	2	1.94	0.90	190
4. Gender	1	6	1.05	0.91	49
5. Current room	2	2	2.00	0.00	27
6. Current bed	1	2	1.80	0.40	100
7. Under Dr. Care	1	2	1.50	0.50	30
8. Under Dr. Care	1	2	1.50	0.50	30
9. Admission reason	1	3	1.33	0.82	175
10. Allergy	3	1	2.6	1	3
11. Chronic condition	1	7	12.41	1.70	29
12. Disease/ problem history	1	10	2.19	2.72	42
13. Care plan	4	9	2.6	1.23	13
14. Status	1	1	1.00	0.00	181
15. Contraption	1	1	1.00	0.00	11

Clinical Narratives

Patient is a 80 y/o female with hx of CAD, DM, HTN, left PICA stroke who presented to the ED after a fall. She was admitted after being found to be demented and unclear if this was a new or old diagnosis. Basic labs for dementia were sent including TSH, B12, folate which were normal. She has a history of angiodysplasia and old PICA infarct. Like a patient with Alzheimer's Disease. She was started on donepezil and Quetiapine. PT/OT evaluated her and felt that she was unable to be d/c home with surveillance.



Overall discussion

The background of the slide is white with abstract blue geometric shapes on the right side. These shapes include overlapping triangles and polygons in various shades of blue, from light to dark, creating a modern, layered effect.

Round-up

The background features abstract, overlapping geometric shapes in various shades of blue, ranging from light sky blue to deep navy blue. These shapes are primarily located on the right side of the slide, creating a modern, layered effect. The rest of the slide is a plain white background.