Ontological Spring
Naumburg, Germany - April 17-20, 2002

Ontologies, Terminologies and Knowledge Bases in the Biomedical Domain

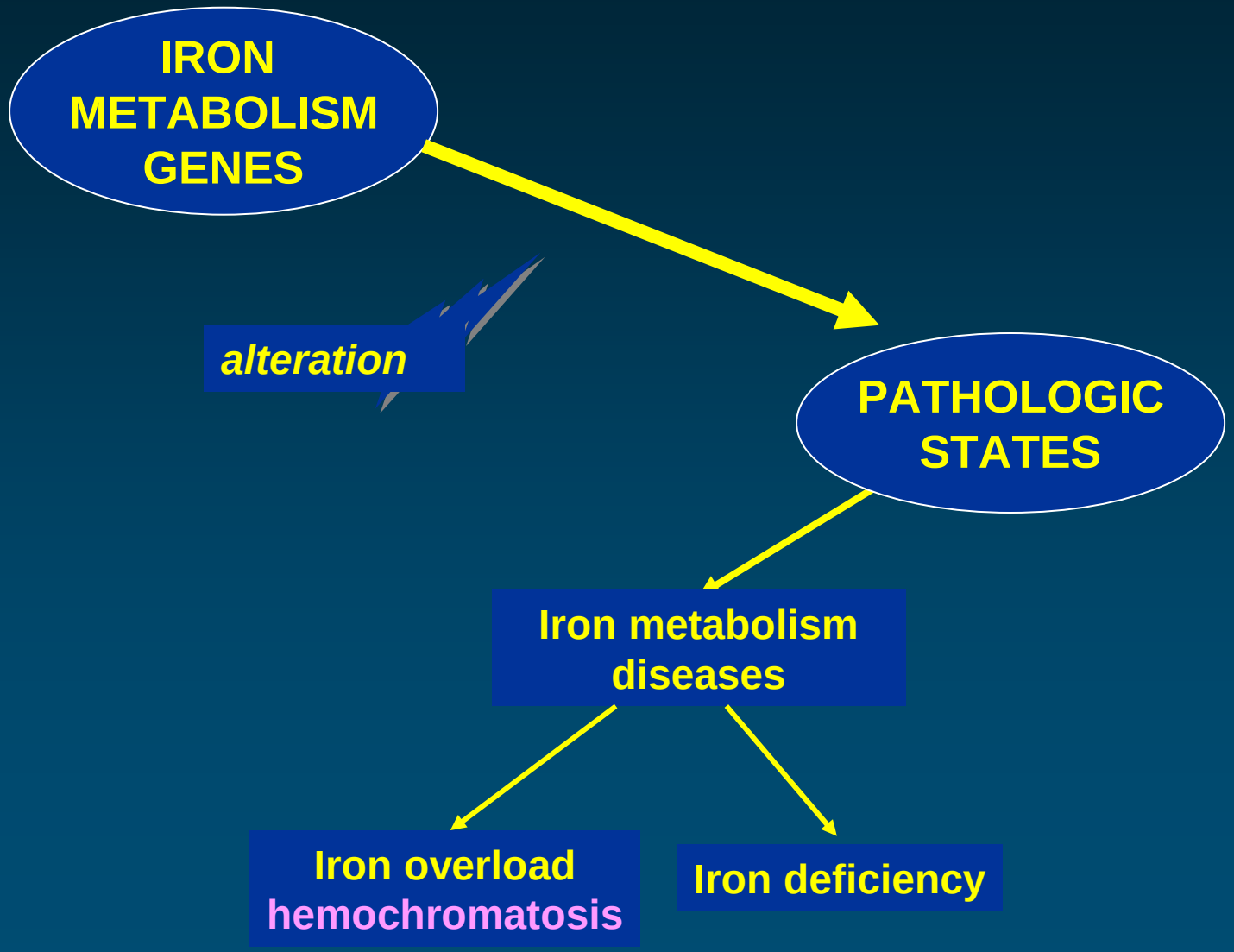
Anita Burgun
Medical School / Univ. Hospital
Rennes, France

Olivier Bodenreider
National Library of Medicine
Bethesda, Maryland - USA

Outline

- Case study: hemochromatosis
- Terminologies vs. ontologies
- Knowledge bases vs. ontologies

Hemochromatosis



Hemochromatosis

- Hemochromatosis is a disorder that causes the body to absorb and store too much iron.
- In the body, iron becomes part of hemoglobin, a molecule that transports oxygen in the blood.
- Healthy people usually absorb about 10 percent of the iron contained in the food they eat. People with hemochromatosis absorb about 20 percent. The body has no natural way to rid itself of excess iron, so extra iron is stored in body tissues, especially the liver, heart, and pancreas

Source : www.niddk.nih.gov

Hemochromatosis

- Iron accumulates in body tissues, which may lead to:
  - Joint pain and arthritis.
  - Liver disease, including cirrhosis, cancer, and liver failure.
  - Heart abnormalities, such as congestive heart failure.
  - Impotence.
  - Abnormal pigmentation of the skin, making it look gray or bronze.
  - Damage to the pancreas, possibly causing diabetes.
- Fatigue is frequent.
- Symptoms tend to occur in men between the ages of 30 and 50 and in women over age 50. However, many people have no symptoms when they are diagnosed.

Source : www.niddk.nih.gov

Hemochromatosis

- Genetic hemochromatosis is mainly associated with a defect in a gene called HFE, which regulates the amount of iron absorbed from food. Two mutations, named C282Y and H63D, can cause hemochromatosis. The genetic defect is present at birth, but symptoms rarely appear before adulthood.
- Hemochromatosis may be acquired, e.g., it may be the result of blood transfusions.

Source : www.niddk.nih.gov

Terminologies vs. ontologies

Terminologies

- Information
  - Represented for a given purpose
    - ICD: Classification
    - MeSH: Information retrieval
  - Useful in a given context
    rather than valid universally
  - Does not necessarily support reasoning (inference)

Terminologies

- Classification : organization of things within categories
  - Taxonomy : kingdom, division, class, order, family, genus, species
- Nomenclature: a system of names for things
  - Rules
    - ranks and terminations:
      - division or phylum: -ophyta
      - class: -opsida
      - subclass: -idea
  - Formula
    - Genus name+ specific modifier= species name

Terminologies

- Nosology: classification of diseases
- Terminologies
  - May include guidelines for describing things
  - Terms for parts
- Controlled vocabularies
  - E.g., subject trees
  - Thesauri

Nosology

- Classificatory systems in the eighteenth century
  - the nosologists tried to do for diseases what the botanists had done for plants: find the "natural" divisions which obtained among diseases, discover the real essence, and embody this essence in a suitable definition
- Thomas Sydenham (1624-1689)
  - « It is necessary that all diseases be reduced to definite and cerútain species . . . with the same care which we see exhibited by botanists in their phytologies ».
- Linnaeus (1707-1778)
  - created classification systems to name animals by genus and species and introduced a plant classification. He also created a medical classification *Genera morborum.*

Nosology

- François Bossier de Lacroix (Sauvages) (1706-77)
  - *Nosologia Methodica*
  - conceived of nosology as a practical discipline providing practitioners with a compass to chart their voyages through the complex sea of symptoms.
- William Cullen
  - Nosology is the key for an improvement of therapeutics
  - The absence of a classification system « would make the study of physic absolutely impossible for if we cannot arrive at some distinction of diseases, we must act at random ».

Terminologies

- Practical purposes
- Reporting on causes of death
  - Child mortality London
  - Plague London
- Nineteenth century
  - William Farr
  - Jacques Bertillon

Terminologies

- Needs for reporting on disease cases (WHO)

Outbreak News - February

Ebola haemorrhagic fever in Gabon - Update 18

Ebola haemorrhagic fever in Gabon - Update 17

Meningococcal disease in Ethiopia - Update

Plague in India

Meningococcal disease in Democratic Republic of Congo - Update 2

Leishmaniasis in Pakistan - Update

Isolation of influenza A(H5) viruses in poultry in Hong Kong Special Administrative Region of China

Ebola haemorrhagic fever in Gabon - Update 16

Tularemia in Kosovo - Update 2

Terminologies

- International Classification of Diseases (ICD)
  - Statistical classification
  - Dated back to the 18th century
  - Early revisions had been concerned only with causes of death
  - Scope extended in 1948 to include non-fatal diseases
  - World Health Organization

Terminologies

- E83 Disorders of mineral metabolism

  Excludes dietary mineral deficiency

  - E83.0 Disorders of copper metabolism
  - E83.1 Disorders of iron metabolism

    Haemochromatosis

    Excludes Iron deficiency anaemia

  - …..
  - E83.8 Other disorders of mineral metabolism
  - E83.9 Disorder of mineral metabolism, unspecified
- E87 Nutritional and metabolic disorders in diseases classified elsewhere

Terminologies

- Needs for controlled vocabularies usable for indexing and cataloging the biomedical literature
- Thesauri in which links show the relationship between related terms and provide a hierarchical structure that permits searching at various level of specificity (from « narrower » to « broader »)
- Medical Subject Headings

Hemochromatosis in MeSH

- A disorder due to the deposition of hemosiderin in the parenchymal cells, causing tissue damage and dysfunction of the liver, pancreas, heart, and pituitary. Full development of the disease in women is restricted by menstruation, pregnancy, and lower dietary intake of iron. Acquired hemochromatosis may be the result of blood transfusions, excessive dietary iron, or secondary to other disease. Idiopathic or genetic hemochromatosis is an autosomal recessive disorder of metabolism associated with a gene tightly linked to the A locus of the HLA complex on chromosome 6. (From Dorland, 27th ed) (MeSH)

Hemochromatosis in MeSH

- Iron overload
  - Not hierarchical
- Metal Metabolism, inborn errors
  - Relationship between Hemochromatosis and Metal metabolism, inborn errors: “is generally a” rather than “is a kind of”
  - Thesaurus relationship (useful for information retrieval), not taxonomic relationship (required for reasoning)

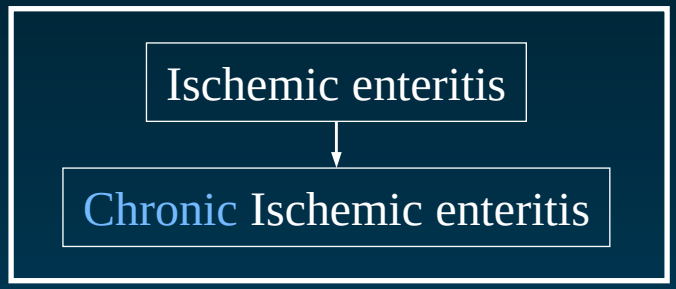
From terminologies to ontologies

- Unified Medical Language System (UMLS)
- More than 50 terminologies
- 2 level structure
  - Semantic Network (134 Semantic Types)
  - Metathesaurus (800,000 concepts)
    - Not all the hierarchical relationships are is-a
    - Assessing consistency through lexical knowledge

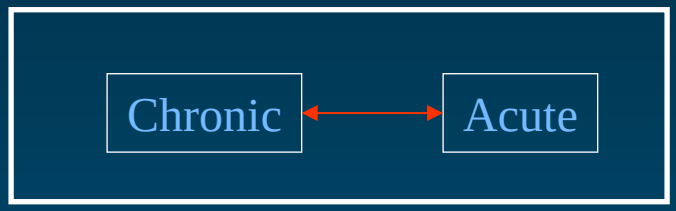
    (Bodenreider, EFMI NLP workshop, Cyprus, Mar 9, 2002)

From terminologies to ontologies

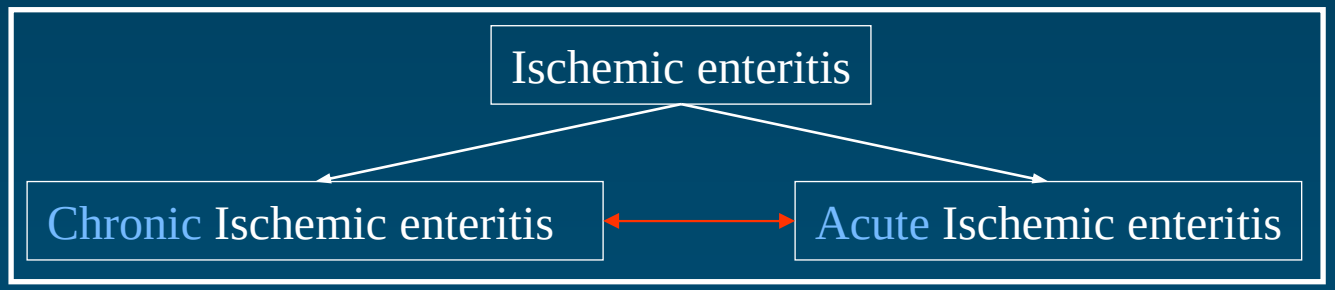
- Adjectival modification generally induces hyponymy



- There are pairs of antonyms



- Consistency
- Siblings
- Opposition



From terminologies to ontologies

- Concept = Cluster of (synonymous?) terms
  - Troisier-Hanot-Chauffard Syndrome
  - von Recklinghausen-Appelbaum disease

  ] Eponym

  - Bronzed cirrhosis
  - Bronzed diabetes
  - Pigmentary cirrhosis of liver

  ] Symptoms
  Complications

  - Hemochromatosis, NEC
  - Hemochromatosis, NOS

  ] Classif. specific
  Underspecified

  - iron accumulation disorders
  - Iron storage disease

  ] Hypernyms
  Microrelations?

From terminologies to ontologies

- Efforts to turn the UMLS into an ontology
  - Medical Ontology Research (MOR) project (NLM) O. Bodenreider
  - S. Schulz, U. Hahn. Medical knowledge reengineering--converting major portions of the UMLS into a terminological knowledge base. Int J Med Inf 2001 Dec;64(2-3):207-21
  - E.g., eliminate terminological cycles

Remaining issues : Standardizing terms

- HUGO Human Gene Nomenclature
- Objectives
  - Simplified information retrieval
  - Database coordination
  - Efficient use of resources
- Meaningful names and systematic polysemy
  - Gene names should be brief and specific and should convey the character or function of the gene.
  - Gene symbols are abbreviation/acronyms of gene names, designed by upper-case latin letters or by a combination of upper-case letters and Arabic numerals. The gene symbol allocated to an inherited clinical phenotype may be based on the name of the disorder, e.g. ACH for achondroplasia. It is usual for this symbol to change when the gene product or function is identified.
  - www.gene.ucl.ac.uk/nomenclature/guidelines.html

Ontologies

- Information
  - Represented independently from any particular purpose
  - Expected to serve several purposes, including reasoning
- Organization based on a theory of the domain, not on a model of the application it is designed for
- Focus on concepts, not terms

Knowledge bases vs. ontologies

Knowledge bases

- Information
  - Represented for a given purpose
    - Diagnosis
    - Therapy
  - Useful in a given context
    rather than valid universally
  - Support reasoning (inference)
    - Decision-support systems

Knowledge bases

- Clinical decision-support systems
- Specific to a subdomain
  - Rule-based expert systems + probabilities
  - MYCIN
- General
  - Diseases, clinical findings, relationships
    - Quick Medical Reference (QMR)
  - Treatments, drugs
- Molecular Biology

Hemochromatosis in QMR

- Associative relationships
  - Has sign, e.g., Hepatomegaly present
- Similar Disorders (by clinical similarity -> prevalence)
  - Very similar, e.g., Androgen induced jaundice
  - Moderately similar, e.g., Alcoholic hepatitis
- Cirrhosis is a parent for Hemochromatosis
  - Focus on the liver damage
  - Relationship between Hemochromatosis and Cirrhosis: “can be evoked when” rather than “is a kind of”
  - “Problem-solving-oriented” relationship (useful for problem solving), not taxonomic relationship

Hemochromatosis in GO Annotations

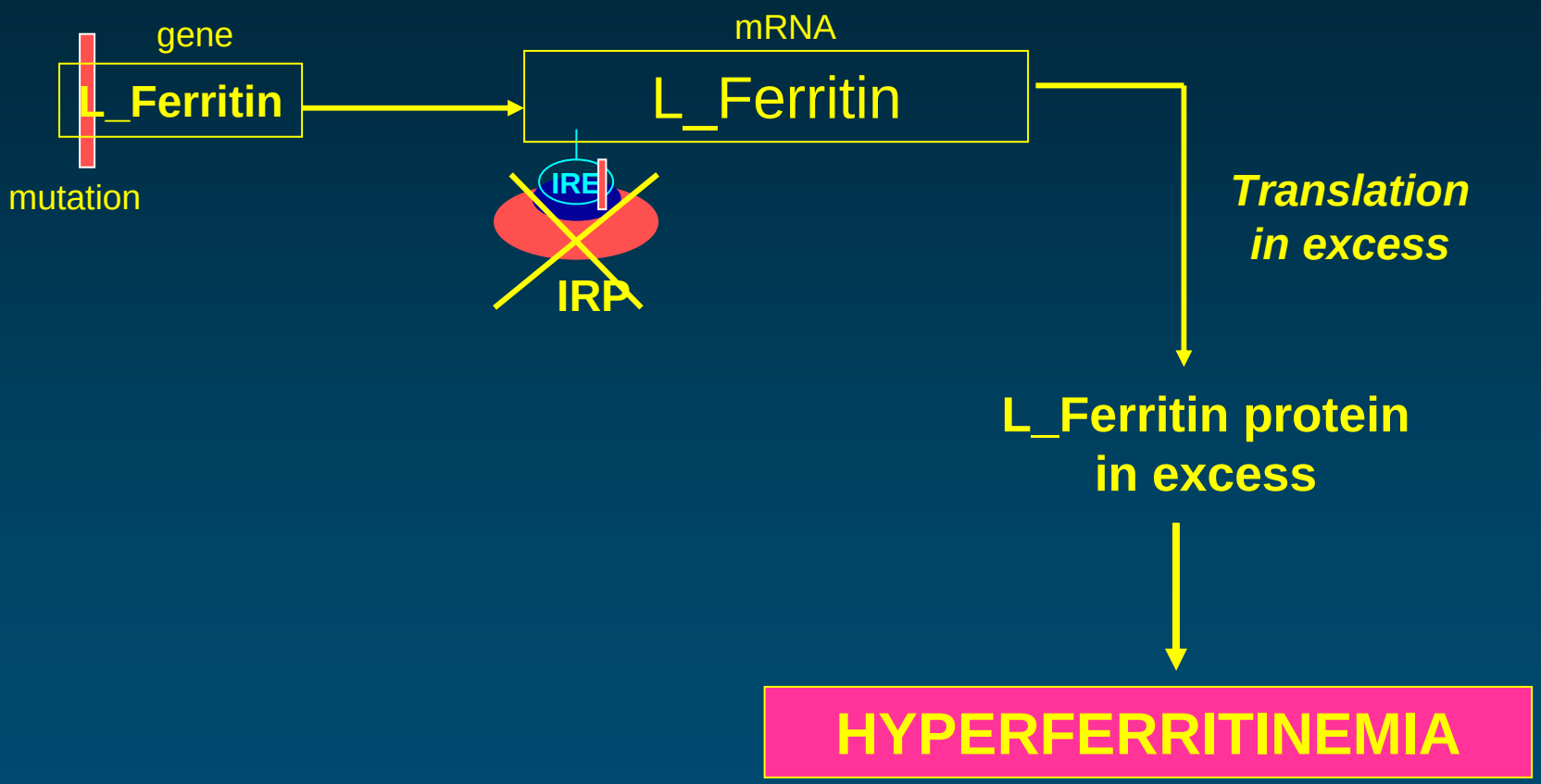
HFE_HUMAN associated with

- GO:0003820 class I major histocompatibility complex antigen
- GO:0005624 membrane fraction
- GO:0005737 cytosol
- GO:0005887 integral plasma membrane protein
- GO:0006461 protein complex assembly
- GO:0006826 iron transport
- GO:0006879 iron homeostasis
- GO:0006898 receptor mediated endocytosis
- GO:0006955 immune response

Hemochromatosis in medical KB

- What is relevant to a particular context
  - Symptom -> diagnosis
  - Genes -> pathway

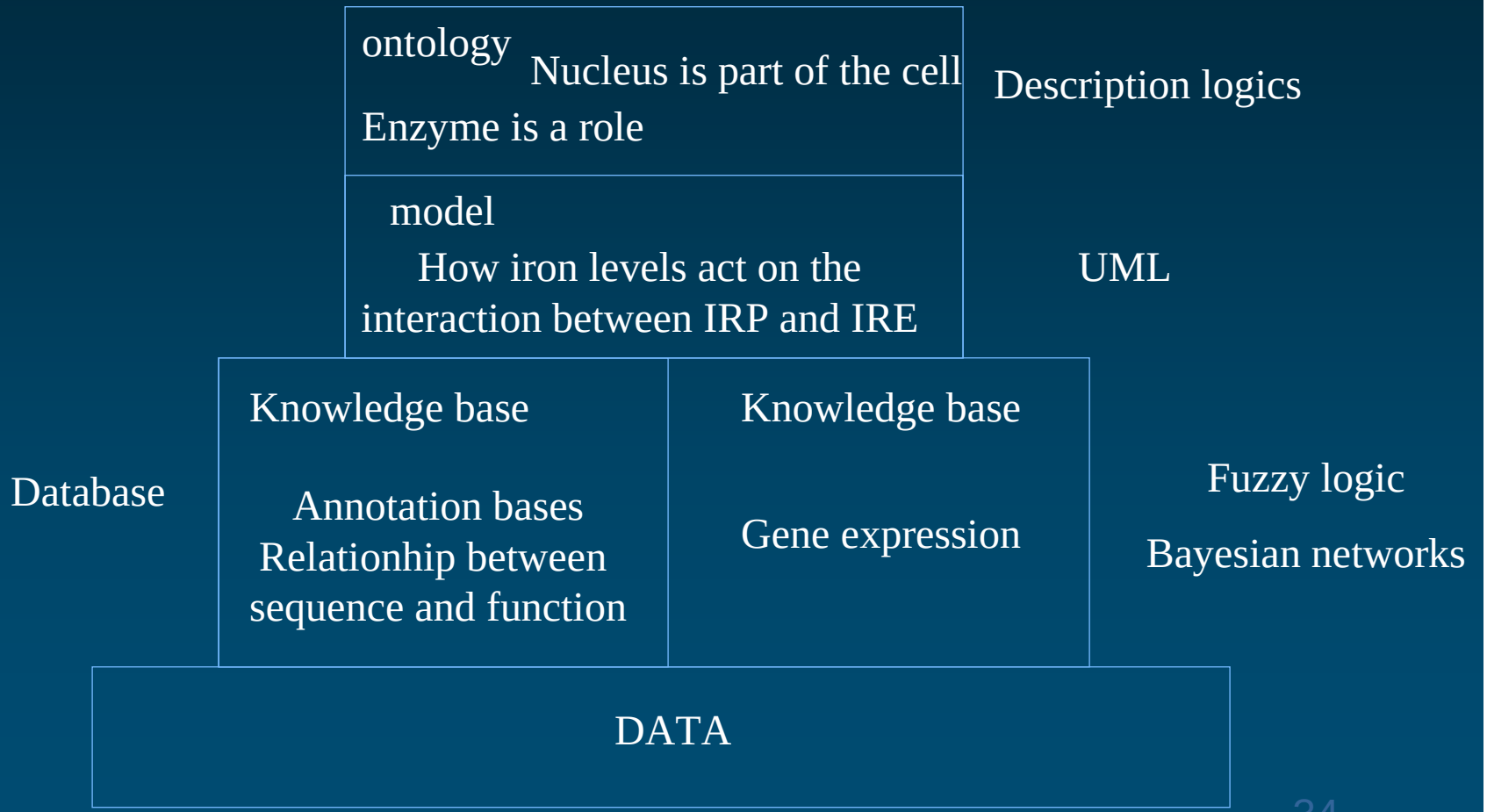
Dynamic models



Remaining issues: organizing KB

- Structured representation of medical information is essential for ensuring the accuracy and reliability of computerized decision support applications.
- As knowledge bases of rules, frames, and other representations grow in size, it becomes increasingly difficult to keep track of what concepts have been represented and of the relationships among them

From KB to ontologies



From KB to ontologies

- Concept = pieces of knowledge that can serve as classes for classifying instances
- Which level of precision?
  - Hepatomegaly present
  - Liver enlarged moderate
- KB as datawarehouse, e.g., genotype and phenotype
- Formal explicit description

Ontologies

- Granularity: coarse grained information
  - Gene Ontology is an ontology for molecular biology and genomics
  - Is not populated with gene products nor gene sequences
- What is always true
- Not facts that are contingently true, being relevant to a particular context

Ontologies

- Provide discrete representation of the world
- Not a dynamic representation of continuous phenomena

Reference

- the "medicine of species" see Michel Foucault, *The Birth of the Clinic: An Archeology of Medical Perception*