Classification and Retrieval of Complex Information in Portfolio Analyses

Christopher G. Chute, MD DrPH
Professor Biomedical Informatics
Vice-Chair, Data Governance
Mayo Clinic College of Medicine

Chair, International Classification of Disease Revision, WHO
Chair, ISO Technical Committee 215 on Health Informatics

NIH Portfolio Analysis Workshop
Bethesda, 23 July 2012

Outline
Classification and Information Retrieval

- Historical and contextual notions
- Major Moving Parts
- Sub-languages and contexts
- Multi-axial classification (multi-dimensional)

The Basic Challenge
Classification

- Starting with unstructured or semi-structured information (documents)
- Assign each document to one or more *categories*
- Strive to
  - Minimize false assignment (precision) – get right
  - Minimize failed assignment (recall) – get all

Natural and *Political*

OBSERVATIONS

Mentioned in a following INDEX,

and made upon the

Bills of Mortality.

1910
1665
245.

BY

Capt. *JOHN GRAUNT*,
Fellow of the *Royal Society*.

With reference to the *Government*, *Religion*, *Trade*, *Growth*, *Air*, *Diseases*, and the several Changes of the said CITY.

—— *Non, me ut miretur Turba, laboro,*
Contentus paucis Lectoribus. ——

The Third EDITION,
much Enlarged.

LONDON,
Printed by *John Martyn*, and *James Allestry*,
Printers to the *Royal Society*, and are to be sold at the
sign of the *Bell* in St. *Pauls* Church-yard.
MDCLXV.

The Table of CASUALTIES.

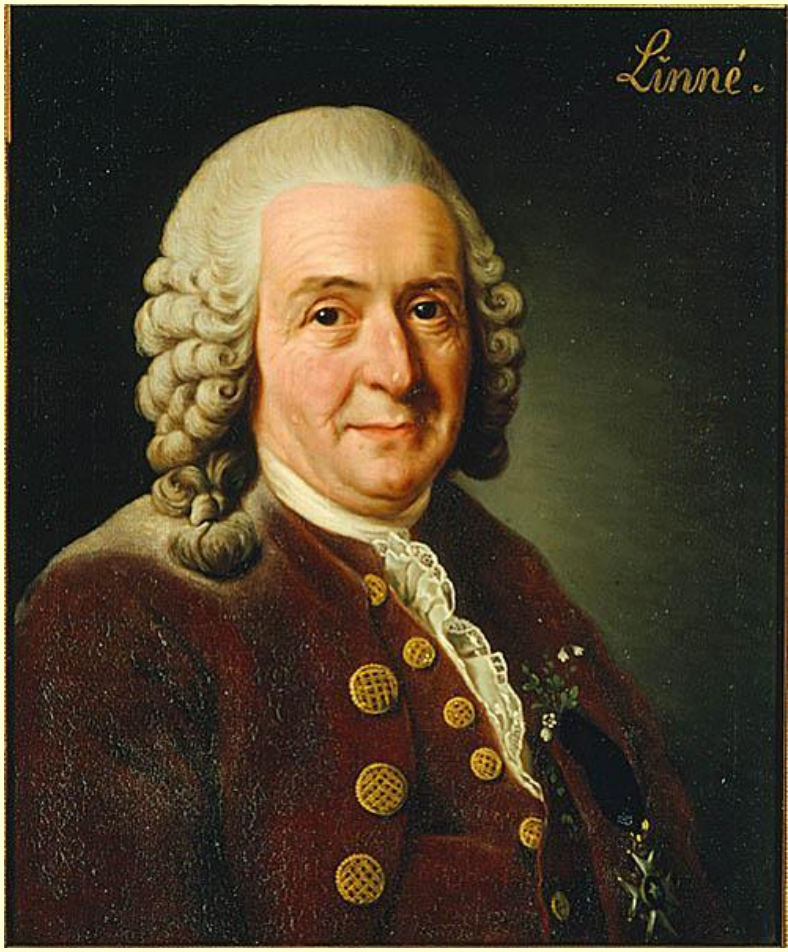
The Years of our Lord	1647	1648	1649	1650	1651	1652	1653	1654	1655	1656	1657	1658	1659	1660	1629	1630	1631	1
Abortive and Stil-born	335	329	327	351	389	381	384	433	483	419	463	467	421	544	499	439	410	4
Aged	916	835	889	696	780	834	864	974	743	892	869	1176	909	1095	579	712	661	6
Ague and Fever	1260	884	751	970	1038	1212	282	1371	689	875	999	1800	2303	2148	956	1091	1115	11
Apoplex and Suddenly	68	74	64	74	106	111	118	86	92	102	113	138	91	67	22	36		
Bleach			1	3	7	2				1								
Blasted	4	1			6	6			4		5	5	3	8	13	8	10	
Bleeding	3	2	5	1	3	4	3	2	7	3	5	4	7	2	5	2	5	
Bloody Flux, Scouring and Flux	155	176	802	289	833	762	200	386	168	368	362	233	346	251	449	438	352	3
Burnt and Scalded	3	6	10	5	11	8	5	7	10	5	7	4	6	6	3	10	7	
Calenture	1			1		2	1	1			3							
Cancer, Gangrene and Fistula	26	29	31	19	31	53	36	37	73	31	24	35	63	52	20	14	23	
Wolf				8														
Canker, Sore-mouth and Thrush	66	28	54	42	68	51	53	72	44	81	19	27	73	68	6	4	4	
Child-bed	161	106	114	117	206	213	158	192	177	201	236	225	226	194	150	157	112	1
Chrisoms and Infants	1369	1254	1065	990	1237	1280	1050	1343	1089	1393	1162	1144	858	1123	2596	2378	2035	22
Colick and Wind	103	71	85	82	76	102	80	101	85	120	113	179	116	167	48	57		
Cold and Cough							41	36	21	58	30	31	33	24	10	58	51	
Consumption and Cough	2423	2200	2388	1988	2350	2410	2286	2868	2606	3184	2757	3610	2982	3414	1827	1910	1713	17
Convulsion	684	491	530	493	569	653	606	828	702	1027	807	841	742	1031	52	87	18	2
Cramp			1														1	
Cut of the Stone		2	1	3		1	1	2	4	1	3	5	6	4				
Dropsie and Tympany	185	434	421	508	444	556	617	704	660	706	631	911	646	872	235	252	279	2
Drowned	47	40	30	27	49	50	53	30	43	49	63	60	57	48	43	33	29	
Excessive drinking			2															
Executed	8	17	29	43	24	12	19	21	19	22	20	18	7	18	19	13	12	
Fainted in a Bath					1													
Falling-Sickness	3	2	2	3		3	4	1	4	3	1		4	5	3	10	7	
Flox and small Pox	139	400	1190	184	525	1279	139	812	1294	823	835	409	1523	354	72	40	58	5
Found dead in the Streets	6	6	9	8	7	9	14	4	3	4	9	11	2	6	18	33	20	
French-Pox	18	29	15	18	21	20	20	20	29	23	25	53	51	31	17	12	12	
Frighted	4	4	1		3		2		1	1				9	1			

Weights and Measures
William Farr

"The nomenclature is of as much importance in this department of inquiry, as weights and measures in the physical sciences, and should be settled without delay."

- First Annual Report of the Registrar-General of Births, Deaths, and Marriages in England. London: 1839 p. 99.

Flawed Information Model



- Carolus Linnaeus
  Carl von Linné
  - *Genera Morborum* (1763)
- Underscored Content Difficulty
  - Pathophysiology vs Manifestation
    e.g. Rabies as psychiatric disease
- Lacked the Germ Theory of Disease
  - Was not incorporated into an information model

The Continuum Of Biomedical Informatics Bioinformatics meets Medical Informatics

Chasm of Semantic Despair

Chris's Classification Project This Week
International Classification of Disease – 11th Rev

ICD-11
International Classification of Diseases
Eleventh Revision

World Health Organization

- Scientific consensus of ***clinical phenotype***
- Public Health Surveillance
  - Mortality
  - Public Health Morbidity
- Clinical data aggregation
  - Metrics of clinical activity
  - Quality management
    - Patient Safety
  - Financial administration
    - Case mix
    - Resource allocation

Classification Methodologies

- Categorization of things (unprincipled)
- Aristotelian differentia – 4th C BCE
- Syllogisms and predicate logic – 13th Century
- Hierarchical assignment – Linnaeus 18th C
- Information Retrieval theory – Salton 1960s
- Natural Language Processing – Sager 1960s
- Frame Logic – Minsky 1970's
- Description Logics – Horrocks 1990s
  - Ontologies -

Salton's Principles for Information Retreival

- Term Frequency
- Inverse Document Frequency
- Vector space model
  - Relevance ranking of documents by similarity
  - Cosine between the term vector and the document vector

- Led to Latent Semantic Analyses
- de facto classification technique

Term or Concept?
The joys of thesauri

- Lexical strings can have semantic equivalence
  - Renal Cancer
  - Carcinoma of Kidney

- Introduction of “concept-based” indexing
- Rise of meta-thesauri
  - NLM/UMLS Metathesaurus
  - NCI Enterprise Vocabulary System metathesaurus
  - Enumerate synonyms, abbreviations, acronyms, etc.

Application of Concept-based Resources

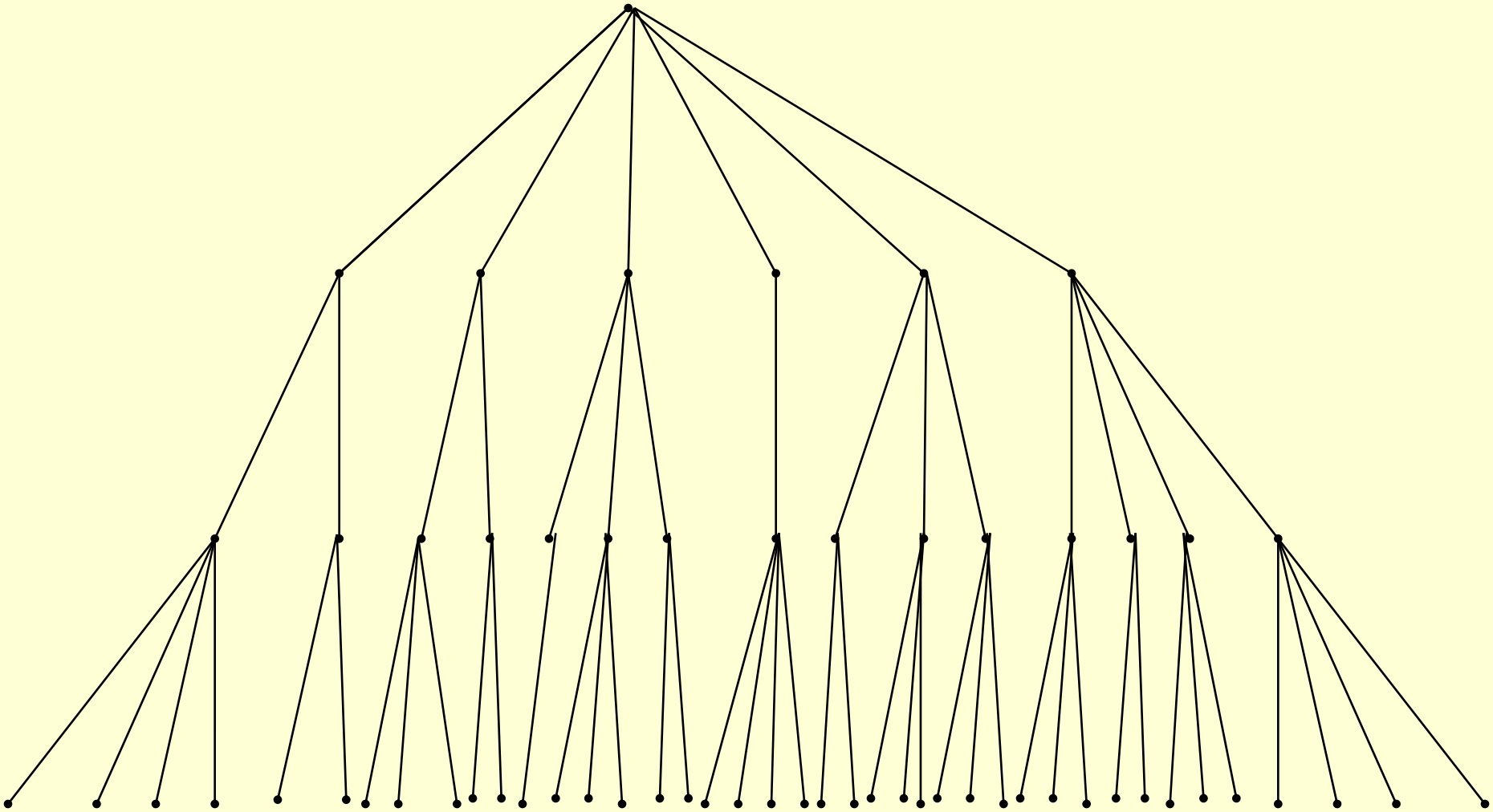
- Variable in Latent Semantic Analyses
  - Assumed that machine-learning techniques can accommodate synonymy
- Variable in Natural Language Processing
  - Function of “named entity” dictionaries

- Always a function of “target” space

Major Moving Parts for any Classification

- The "bins"
  - Categories or target for classification
- Thesauri
  - Alternate names and labels for the bins
- Analytic engine
  - Machinery or methods for "binning" content
  - Logistic regression, neural nets, Bayesian methods
  - Natural Language Processing
  - Rule-based methods (DROOLs, PROLOG, etc)
  - Vector-space models

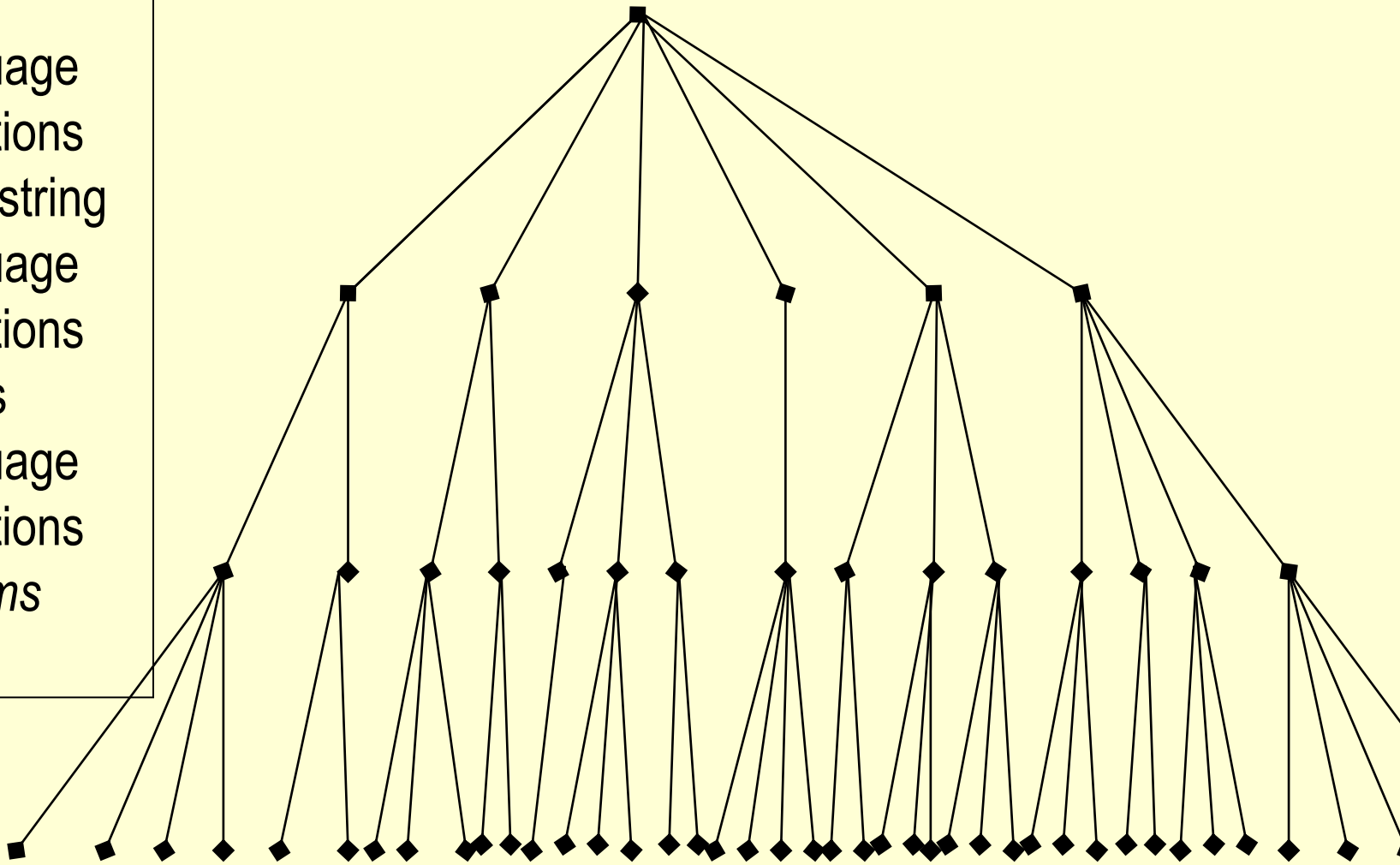
Traditional Hierarchical System
ICD-10 and family



Addition of structured attributes to concepts

Concept name

- Definition
  - Language translations
- Preferred string
  - Language translations
- Synonyms
  - Language translations
- *Index Terms*



THE CONTENT MODEL

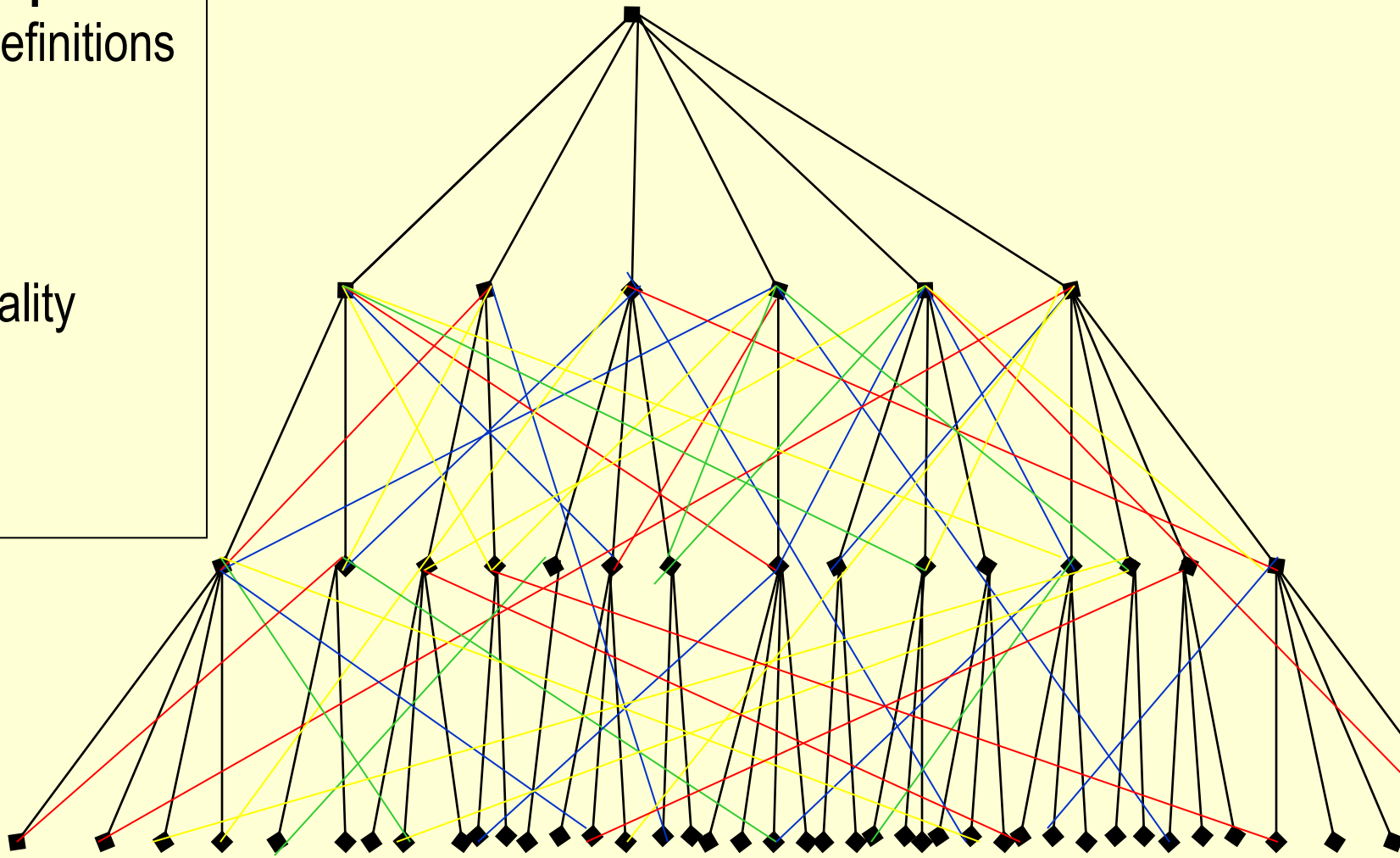
Any Category in ICD is represented by:

1. ICD Concept Title
   - 1.1. Fully Specified Name
   - 1.2 Preferred Name
   - 1.3 Synonyms
2. Classification Properties
   - 2.1. Parents
   - 2.2 Type
   - 2.3. Use and Linearization(s)
3. Textual Definition(s)
4. Terms
   - 4.1. Base Index Terms
   - 4.2. Inclusion Terms
   - 4.3. Exclusions
5. Body System/Structure
   - 5.1. Body System(s)
   - 5.2. Body Part(s) [Anatomical Site(s)]
   - 5.3. Morphological Properties
6. Manifestation Properties
   - 6.1. Signs & Symptoms
   - 6.2. Investigation findings
7. Causal Properties
   - 7.1. Etiology Type
   - 7.2. Causal Properties - Agents
   - 7.3. Causal Properties - Causal Mechanisms
   - 7.4. Genomic Linkages
   - 7.5. Risk Factors
8. Temporal Properties
   - 8.1. Age of Occurrence & Occurrence Frequency
   - 8.2. Development Course/Stage
9. Severity of Subtypes Properties
10. Functioning Properties
    - 10.1. Impact on Activities and Participation
    - 10.2. Contextual factors
    - 10.3. Body functions
11. Specific Condition Properties
    - 11.1 Biological Sex
    - 11.2. Life-Cycle Properties
12. Treatment Properties
13. Diagnostic Criteria

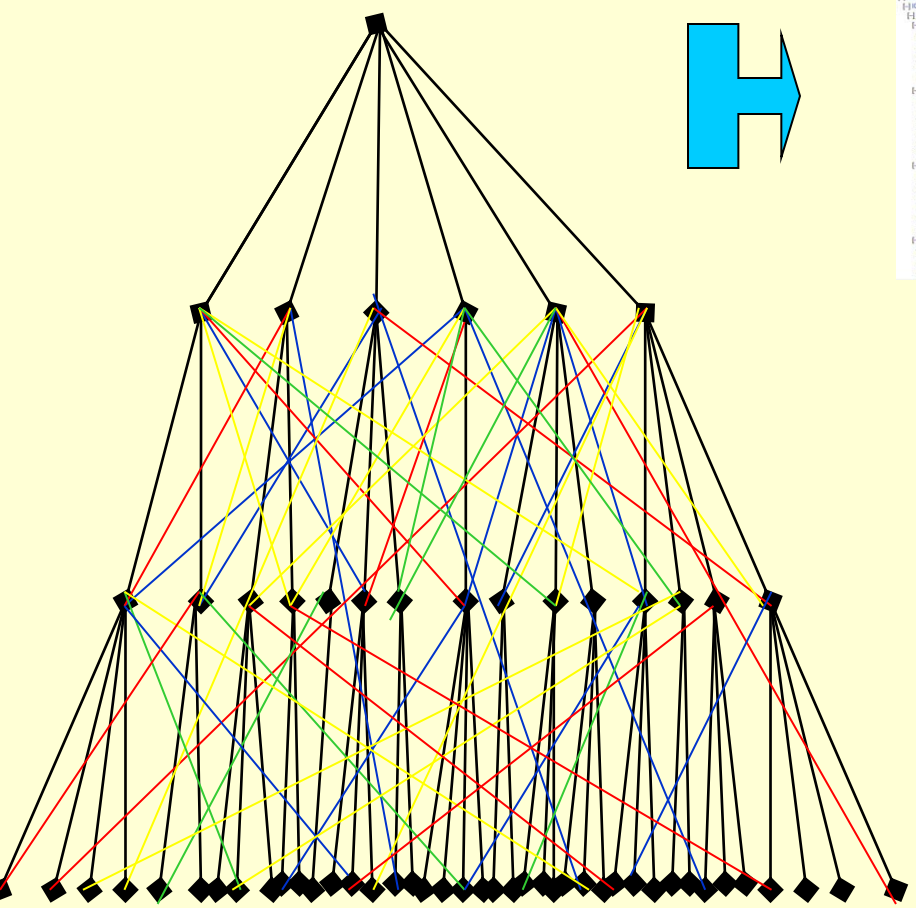
Addition of semantic arcs - Ontology

Relationships

- Logical Definitions
- Etiology
- Genomic
- Location
  - Laterality
- Histology
- Severity
- Acuity

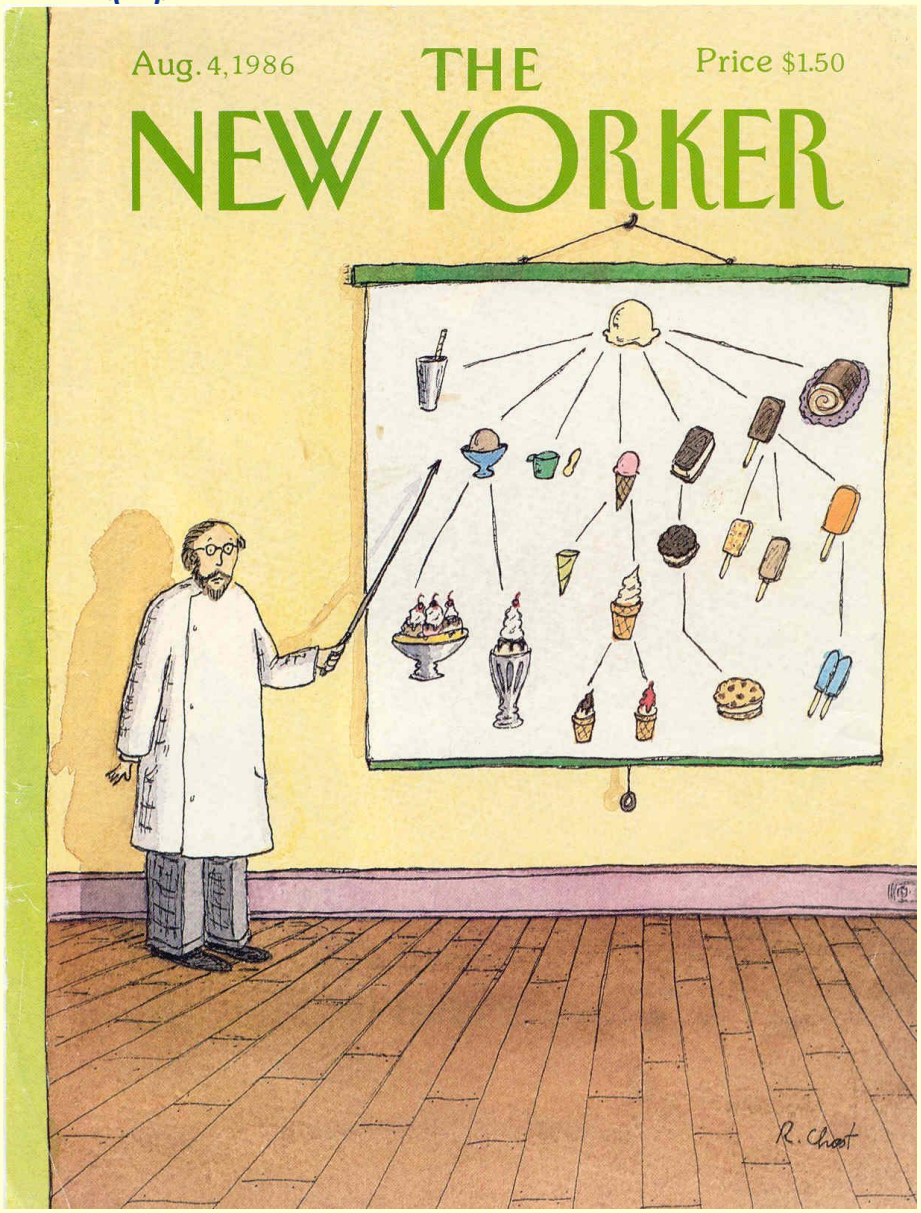


Linear views may serve multiple use-cases
Morbidity, Mortality, Quality, …



Classifications

- Statistical – mutually exclusive and exhaustive
  - Goal is not to count things twice
  - ICD is a classic example – arbitrary hierarchy
- Comprehensive – exhaustive
- Hierarchical
  - Must not create “cycles” in graph tracing
  - Acyclic directed graphs
- Convenience – multiple assignment OK
  - Often not comprehensive



"Clinical" Ontology?

- A common vocabulary
- A shared exchange
  - among people
  - among software agents
  - between people and software
- Facilitates Data Re-use
- Vocabulary standards to allow interoperability

Formal Ontology

- Based upon Description Logic
  - Computable subset of First Order Predicate Logic
  - OWL – Web Ontology Language most familiar
- Concepts
  - Properties and attributes of concepts
  - Constraints on properties and attributes
  - Instances
- The components of ontologies
  - Classes
  - Attributes
  - Relations
  - Axioms

Are there any practical "Formal Ontology"

- May be an oxymoron
- Description Logics cannot accommodate probability – "usually true" relationships
- SNOMED
  - Most concepts still primitive
  - Description Logic definitions – while true, are not useful
  - Myocardial Infraction – death of myocytes
- Gene Ontology

Sub-language Theory In machine translation

- Whole, random document input - ☆☆☆☆
  - Performance is typically laughable
  - Bablefish programs
  - Cannot disambiguate
- Domain constrained ★★☆☆
  - Dramatic improvement in performance
  - Context constrains allowable meaning
- Sub-section training within domain ★★★★
  - Precipitation section of a weather report

Generalization of Sub-language to Healthcare Documents – Clinical Dictations

- Natural Language Processing on random clinical documents ★☆☆☆
- NLP on Discharge Summary ★★☆☆
- NLP on Radiology Report ★★★☆
- NLP on "impression" of radiology report ★★★★

Past medical history, eye exam, family history, …

Generalization of Sub-Language to NIH Portfolio

- Recognize that grants have logical structure
  - Abstract, Aims, Public statements, etc.
- Have much more "implicit" structure
  - Background, significance, previous work, methods

- How much "structure" can be inferred from existing grant submissions?
- How much "structure" can be encouraged/required from future submissions?

Some Observations on the RCDC Research, Condition, and Disease Categorization

- Categories are highly discordant
  - Common conditions - Cardiovascular
  - Rare diseases - Batten Disease
  - Social cultural states – Underage Drinking
  - Research functions - Clinical Trials
  - Domains - Mind and Body
  - Demographic – Pediatric
  - Goals - Youth Violence Prevention
  - Substances – Estrogen
- Challenge of overlapping categories

Topic vs. Impact Problem

- Grants about specific topics
  - specific gene pathway
- Relevant to large domain topics
  - Cancer risk
- Leverage knowledge-bases
  - Association between concepts asserted outside grant
  - Ontologic principles
- Crucial to inclusiveness metrics

Secondary Use of EHR data Themes & Projects

SHARPn.org

Themes			Projects	Players
Data Normalization	Phenotype Recognition	Data Quality and Evaluation Frameworks	Clinical Data Normalization	IBM, Mayo, Utah, Agilex
			Natural Language Processing (NLP)	Harvard, Group Health, IBM, Utah ,Mayo, MIT, SUNY, i2b2, Pittsburgh, Colorado
			High-Throughput Phenotyping	CDISC, Centerphase, Mayo, Utah
			UIMA and Scaling Capacity	IBM, Mayo
			Data Quality	Mayo, Utah
			Evaluation Framework	Agilex, MN HIE, Mayo, Utah

Fully Automated Classification Ready for Prime-Time?

- Most serious researchers would give a qualified no.
- Realistic and practical only under highly specific circumstances (sub-language, narrow domain)
- Methods are still in active research
- While some NLP is commodity, large-scale classification remains unsolved
- NLM is among world leaders
  - Still uses human over-read for MeSH assignment

Advice
Sub-language domains

- Consider a multi-dimensional classification
  - Assignment within homogenous categories
    - Mutually exclusive options within domain – with “None”
  - Diseases, demographics, functions, cultural issues
  - Techniques, substances, topic domains
- Consider increasing sub-structure of submissions
  - Hypothesis, impact, significance, methods,
- Train classifiers specific to each sub-structure and for each dimension
- Engage the NLM in these processes
  - There is still a lot of R&D to make practical tools