Ontology as a tool for automated interpretation

Boris Mirkin
Department of Data Analysis & AI, NRU HSE, Moscow RF
Department of CS, Birkbeck University of London UK

Joint work with
T. Fenner (U of London),
S. Nascimento (NU Lisbon),
E. Chernyak (NRU HSE)

Supported by
-**Research and Academic Funds of NRU HSE: «Teacher-student» 2011-14 and Research Lab Decision Choice and Analysis 2010-pr.;**
-**grant of Portuguese Science and Technology Foundation 2007-2011 (to SN & BM)**

Plenary talk at "The 16th International Conference on Artificial Intelligence: Methodology,

Outline I

- Intro: an outsider bird's eye view at AI history
- Hierarchical ontology / Taxonomy / Relation "…is a …"
- A system for the "vertical" computational interpretation:
  - Granularity levels: Concepts-themes-elements
  - Types of vertical interpretation
    - Annotation
    - Elemental query set and overrepresentation
    - Thematic query set lifted to higher ranks in a taxonomy
  - [Developing a (fuzzy) query set]

Outline 2

- Parsimoniously lifting a thematic query set (PARL)
- Application cases
  - Individual gene histories and LUCA
  - Representing research of organization
  - Analysis of residents' complaints
- Conclusion

Outsider's view of history of AI

I. **Romantic AI: Turing test (1940 – 1960)**

II. **Deductive AI: reasoning to automate (1960 – 1990)**

III. **Inductive AI – data analysis, data mining, knowledge discovery (1990 – …)**

IV. **Synthesis: Ontology (2010 – …)**

Outsider's view of history of AI

I. **Romantic AI: Turing test; perceptron; machine translation (1940 – 1960)**

Turing test: a joke ?

II. **Deductive AI: reasoning to automate (1960 – 1990)**

In spite of Gödel's theorem (?)

III. **Inductive AI – data mining, knowledge discovery (1990 – …)**

Computational Intelligence (2005) versus deductive AI: (1) neural networks, (2) fuzzy sets and logics, (3) genetic and evolutionary algorithms – converging to modeling AI as an evolving phenomenon

IV. **Synthesis: Ontology (2010 –)**

SNOMED CT, GO, ACM CCS…

Example I: SNOMED CT – A set of bio-medical hierarchical ontologies and semantic mappings among them

Systematized Nomenclature of Medicine--Clinical Terms (SNOMED CT)

– a multinational effort in computerization of all things related to health and medicine, **~311 000** terms so far

SNOMED Ontologies: Whats and Whys

"SNOMED CT is

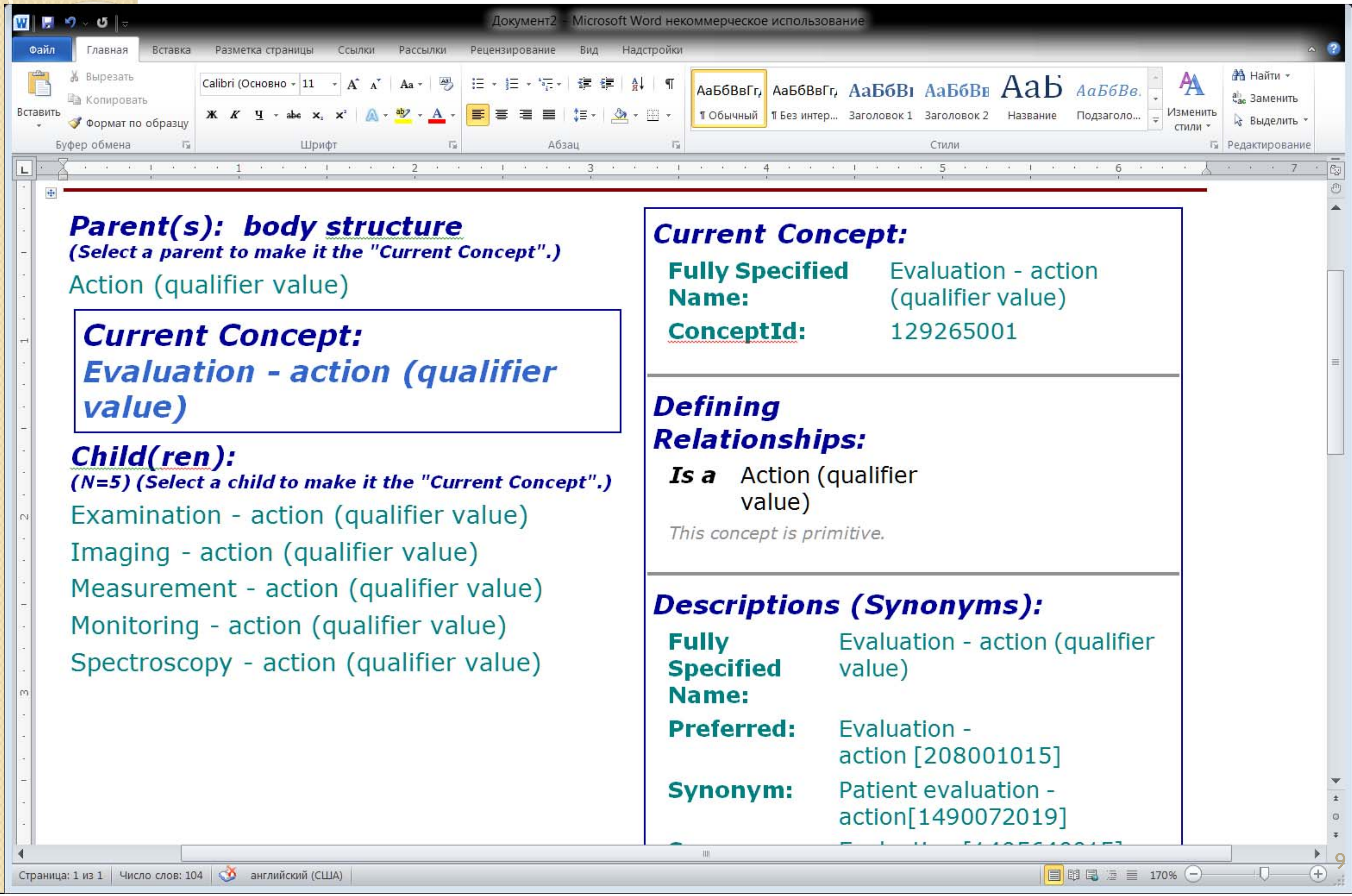
- a clinical healthcare terminology
- a resource with comprehensive, scientifically-validated content essential for electronic health records
- a terminology that can cross-map to other international standards already used in more than fifty countries

SNOMED CT provides the core general terminology for the **electronic health record (EHR)** and contains more than **311,000 active concepts** with unique meanings and formal logic-based definitions organized into hierarchies. When implemented in software applications, SNOMED CT can be used **to represent** clinically relevant information **consistently, reliably and comprehensively** as an integral part of producing electronic health records." **IHTSDO**

Hierarchies in SNOMED CT

Clinical finding/disorder
Procedure/intervention
Observable entity
Body structure
Organism
Substance
Pharmaceutical/biologic product
Specimen
Special concept
Physical object
Physical force
Event
Environment or geographical location
Social context
Staging and scales

Example 1: SNOMED CT – A set of bio-medical hierarchical ontologies (2012)



Example II:

The 2012 ACM Computing Classification System: ACM-CCS-2012

Hierarchical Taxonomy – 5-6 Layers

Example II: ACM-CCS-2012 Taxonomy – Layer One, 14 categories

- General and reference
- Hardware
- Computer systems organization
- Networks
- Software and its engineering
- Theory of computation
- Mathematics of computing
- Information systems
- Security and privacy
- Human-centered computing
- Computing methodologies
- Applied computing
- Social & professional topics
- Proper nouns: People, technologies and companies

Example 2: ACM-CCS Taxonomy – Layer two, Maths of computing

- **Mathematics of computing**
  - Discrete mathematics
  - **Probability and statistics**
    - Statistical paradigms
      - Queueing theory
      - Contingency table analysis
      - Regression analysis
      - Time series analysis
      - Survival analysis
      - Renewal theory
      - Dimensionality reduction
      - **Cluster analysis**
      - Statistical graphics
      - Exploratory data analysis
    - Multivariate statistics

Mathematics of computing (cont.)

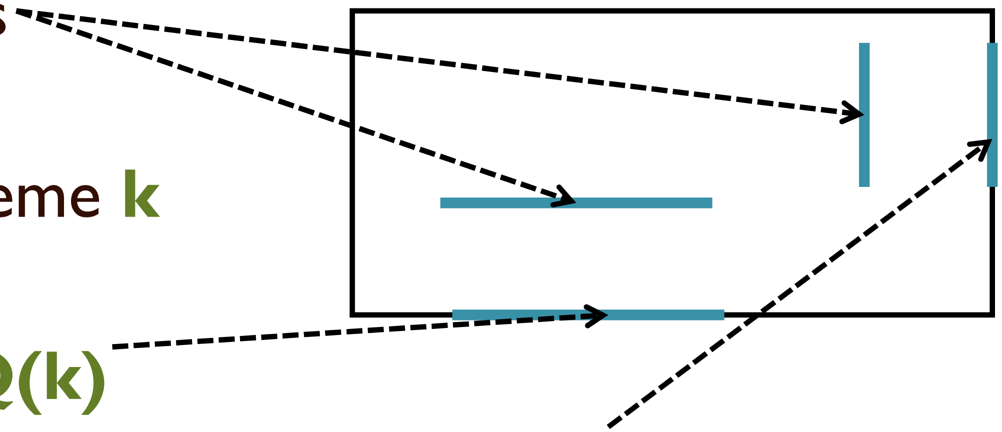
- Mathematical software
- Information theory
- Mathematical analysis
  - Numerical analysis
  - Mathematical optimization
  - Differential equations
  - Calculus
  - Functional analysis
  - Integral equations
  - Nonlinear equations
  - Quadrature
- Continuous mathematics

“Interpretation”: meaning (?)

- To interpret: “to explain or tell the meaning of, that is, present in understandable terms” (Merriam-Webster)
- “Explanation” must be “concise.”
- Generalization: a special case of interpretation (2a) “generalize”: (1) to give a general form to, (2a) **to derive or induce (a general conception or principle) from particulars**, (2b) to draw a general conclusion from (Merriam-Webster)
- Annotation: “a note added by way of comment or explanation”(Merriam-Webster)

Basic Computational Interpretation:

1. **Build Theme-to-Element relevance matrix**, say, KeyPhrase-to-Text or Motif-to-ProteinSeq or ResearchSubject-to-ResearchTeam



2. **Build elemental query sets Q(k) for themes**
3. **Build thematic annotations A(j) for elements**

Interpretation of **thematic query sets I**:

Two types of concepts– **themes, elements**

Concept granularity

Concept

Concept

Concept

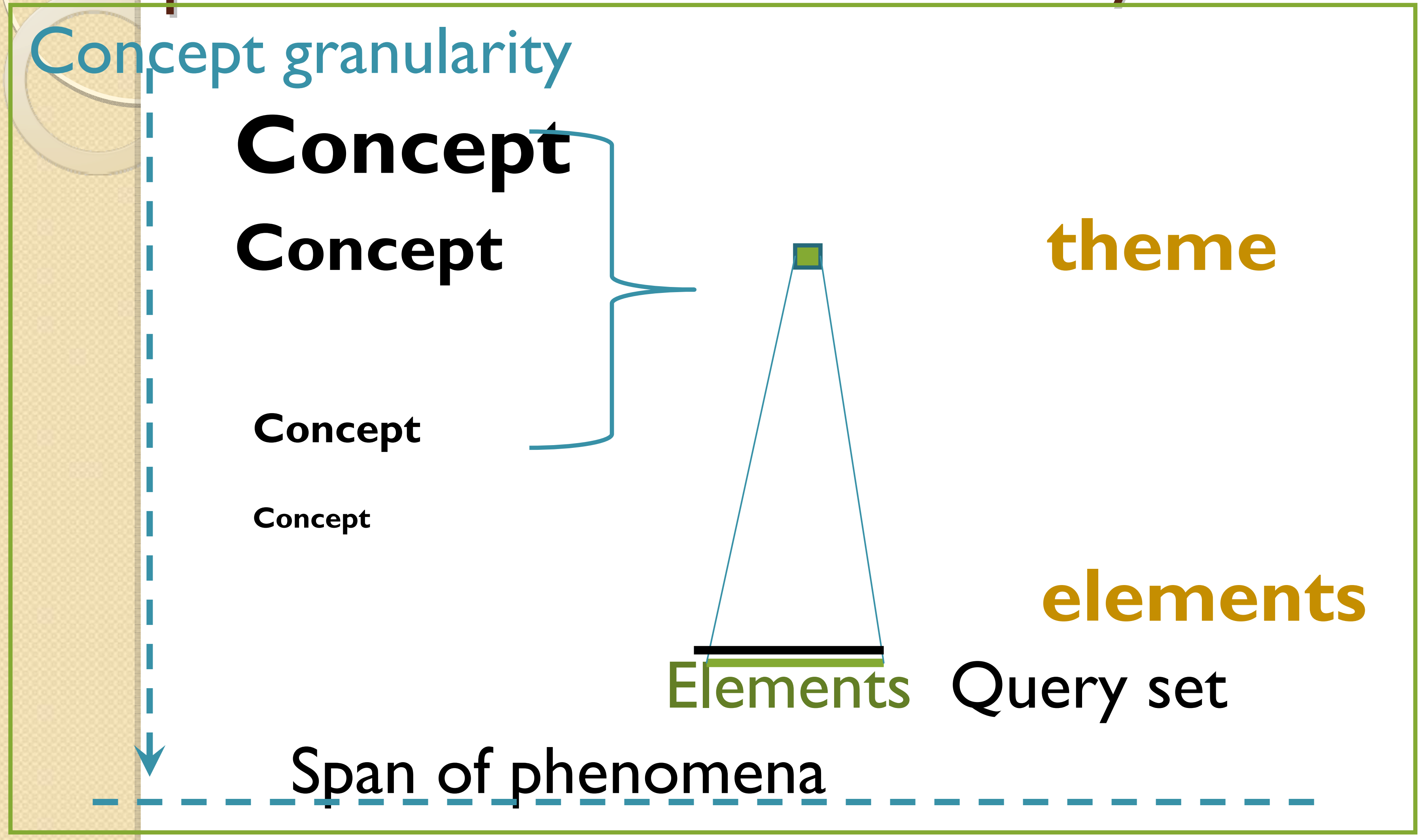
Concept

Taxonomy concept granularity: **themes** (my subject)

Finer granularity: **elements**

Span of phenomena

Interpretation of **concept query sets II**:
Interpretation 1: set of **elements** by a **theme**



Interpretation of **concept query sets III**: Interpretation 1: set of **elements** by a **theme**

Bioinformatics: **Q – co-expressed genes, T– genes of a same function**

T

Q

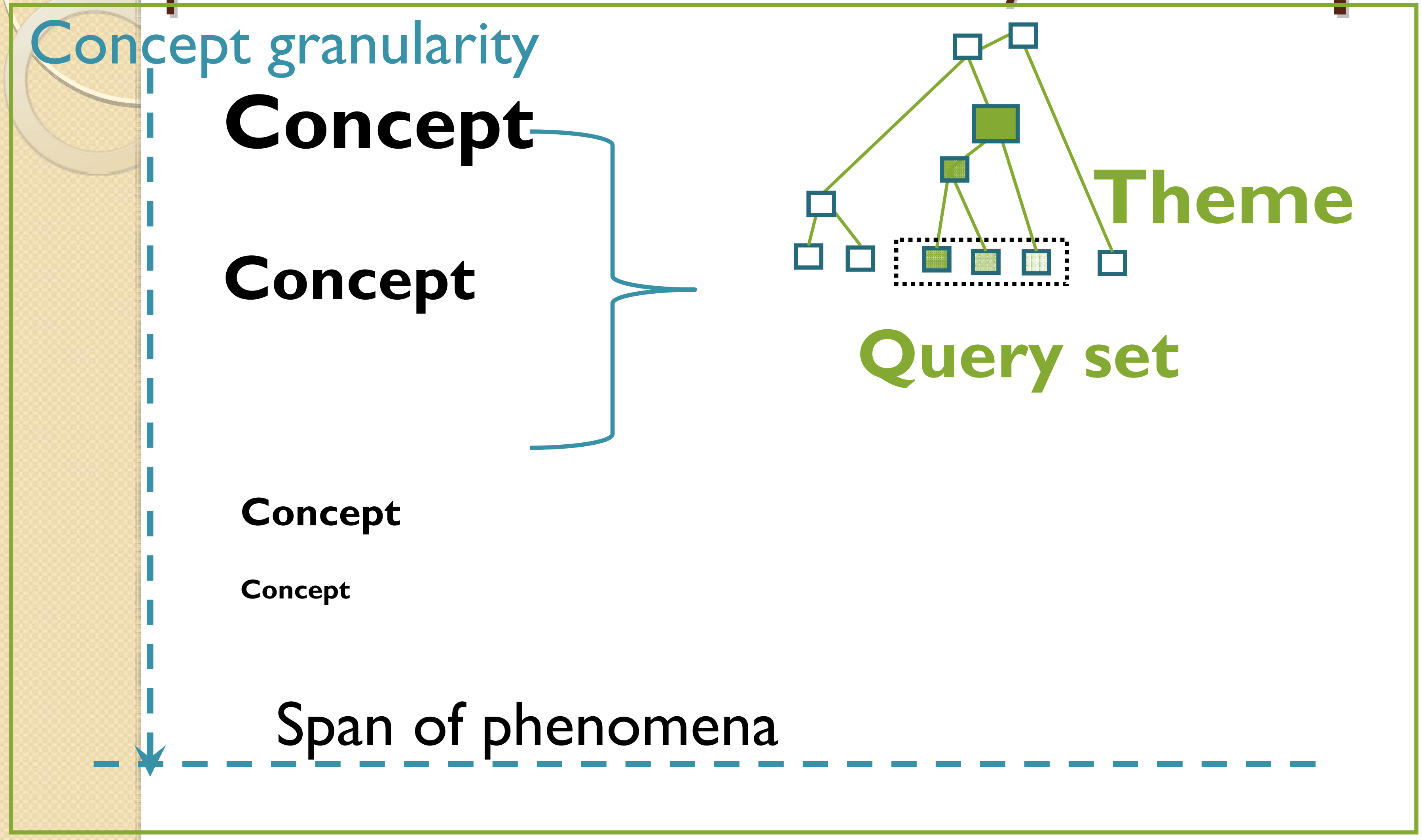
- **Taxonomy concept T**
  Elemental query set Q

Overrepresentation (Robinson 2011)

If Prob(QT/Q) >> Prob(T), annotate Q by concept T

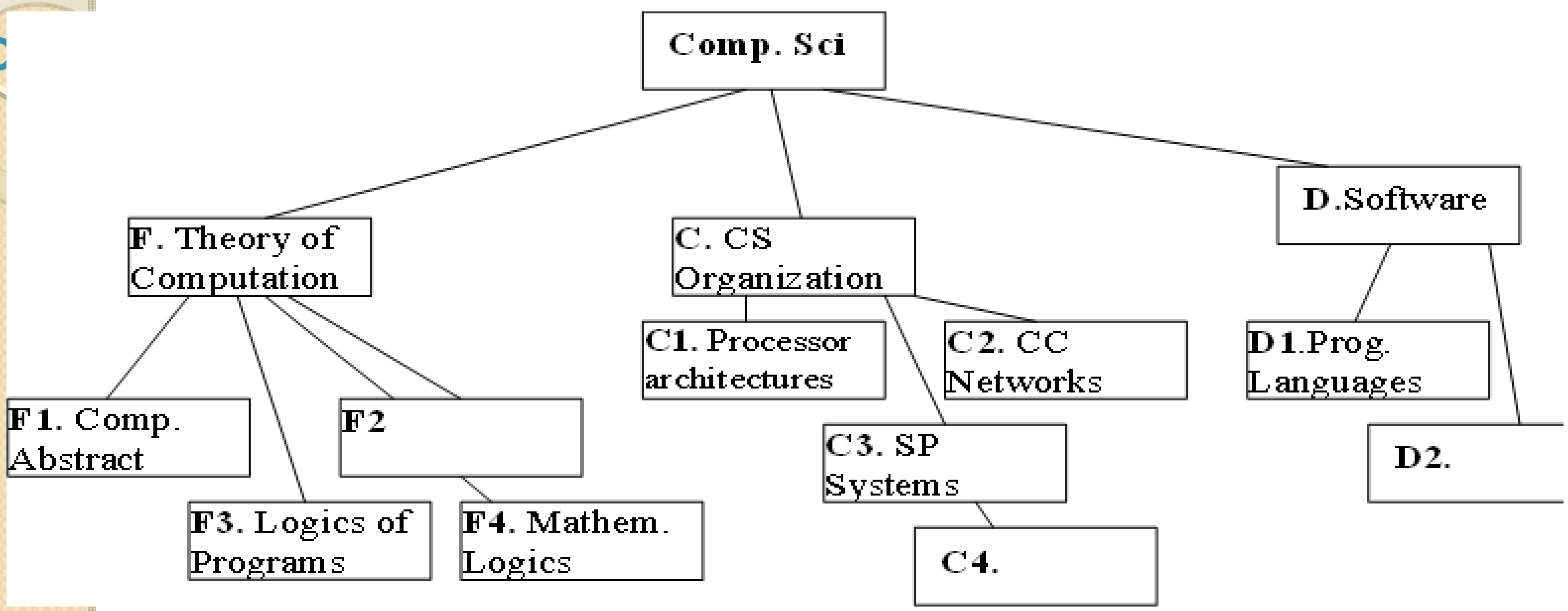
Interpretation of **concept query sets IV**:

Interpretation: set of **themes** by a **Concept**



Interpretation in Domain Taxonomy T T (a)

- Given a T and Out-T-Concept, “intuitionistic p



- Map O-T-Concept as a fuzzy topic set:

F.1 Computation by abstract devices - 0.60

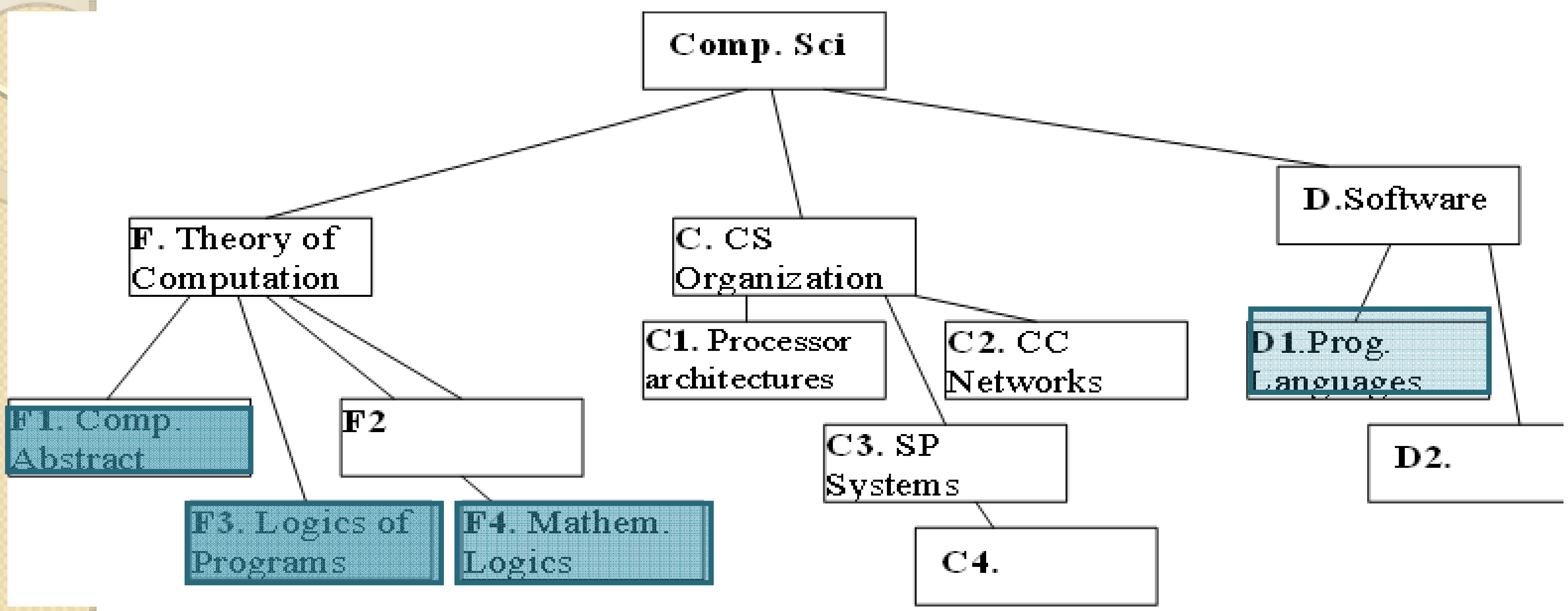
F.3 Logics and meaning of programs - 0.60

F.4 Mathematical logic and formal languages - 0.50

D.1 Programming languages - 0.17. (Euclidean Normed)

Interpretation in Domain Taxonomy I(b)

- Given T and Out-T-Concept **"intuitionistic program."**



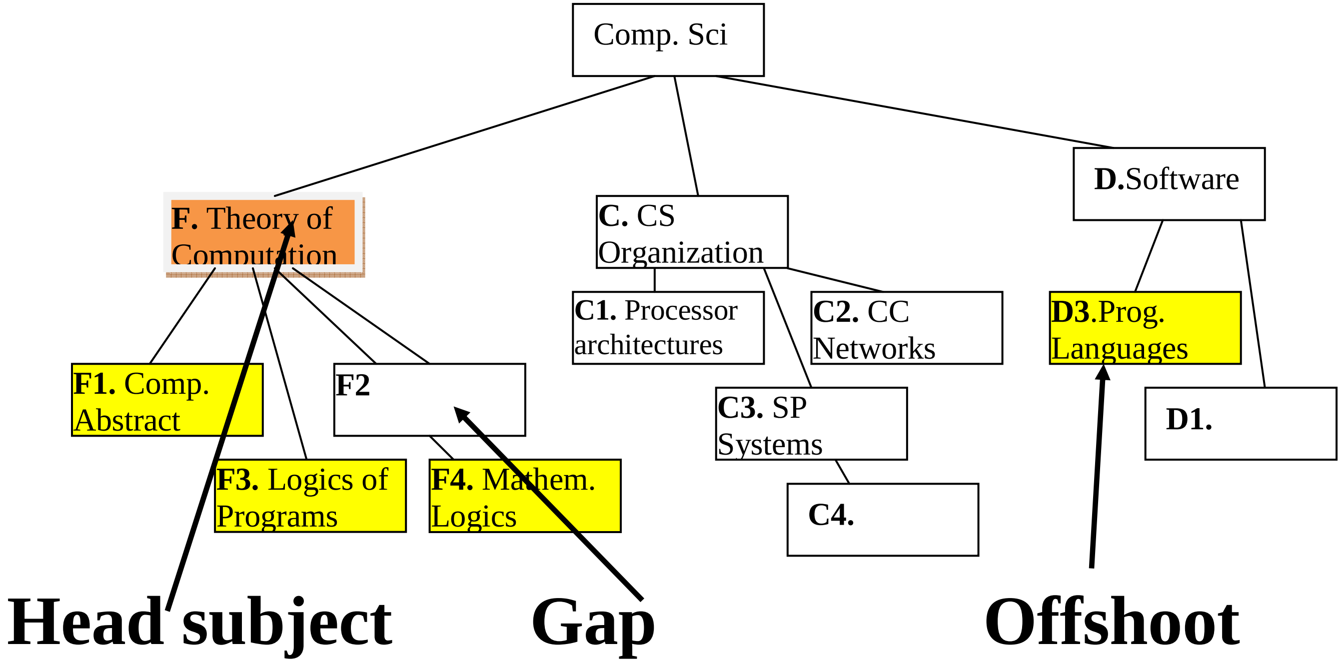
- Map O-T-Concept to Taxonomy as just a fuzzy topic set:

  {F.1 - 0.60, F.3 - 0.60, F.4 - 0.50, D.1 - 0.17} (Euclidean Norm)

Fragmentary

Not cognition friendly

Interpretation in Domain Taxonomy I(c) by Lifting



Interpretation of ***the thematic cluster:***

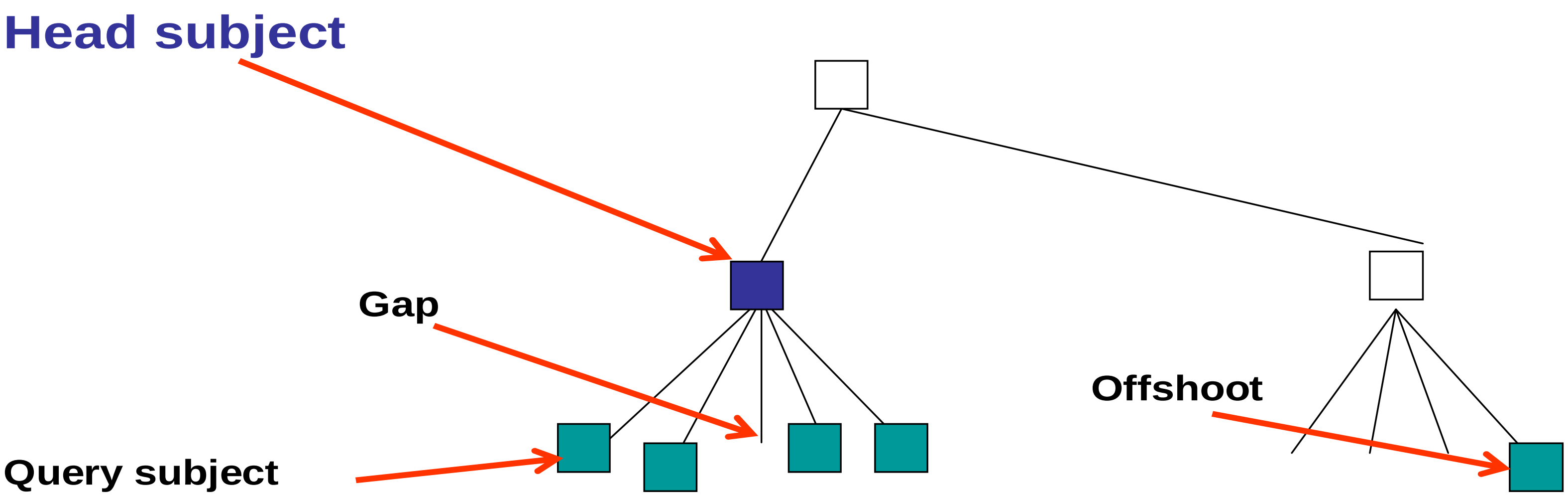
F. Theory of computation

(With a *gap*, F2, and an *offshoot*, D3)

Interpretation of thematic clusters in T by lifting

Lifting penalty function

Represent the thematic clusters in ACM-CCS by higher, more general, nodes depending on the inconsistencies (Lift)



Minimize **H∗#Head_Subj + G∗#Gap + O∗#Offshoot**

15

Criterion balances **the number of Head Subjects** (the higher the ranks, the smaller the numbers) with **those of Gaps/Offshoots** (the opposite)

Algorithmic issues I

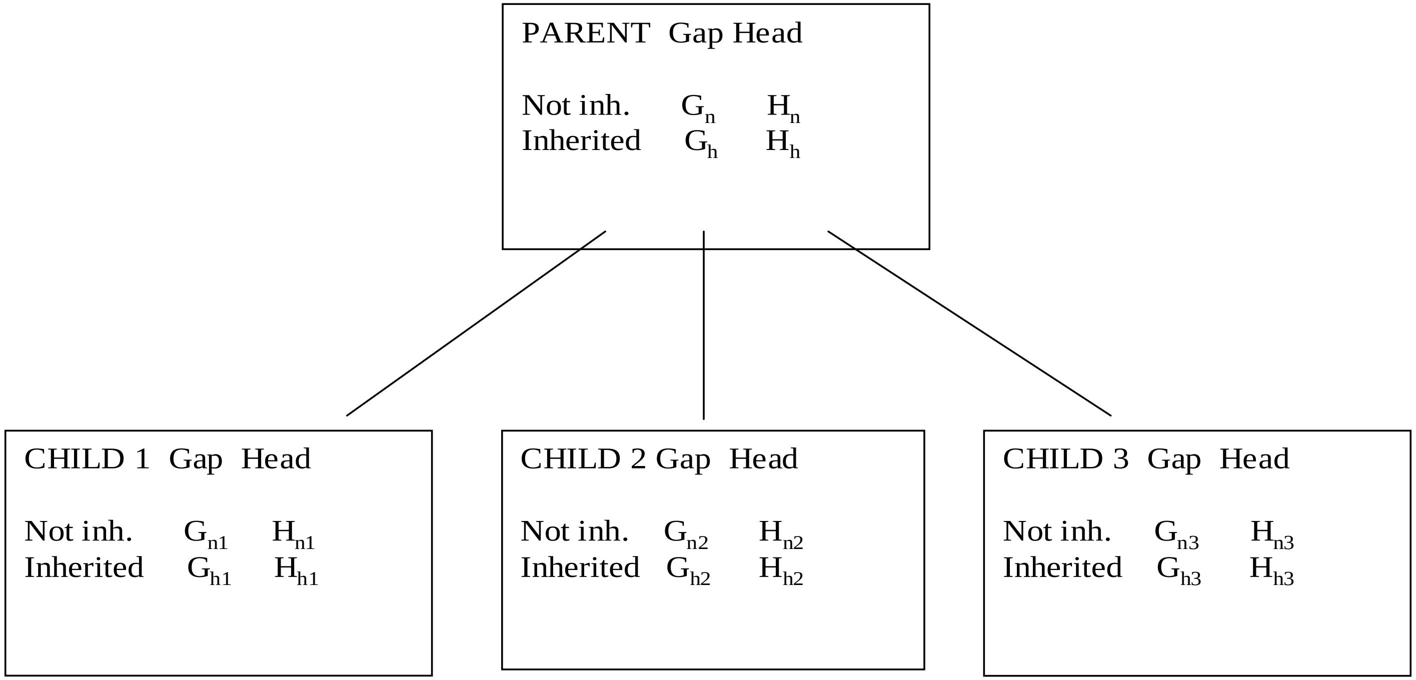
- Cleaning the taxonomy tree of **irrelevant** nodes
- Ways to extend **the fuzzy belongingness values** to all the nodes (no effect on the algorithm but on results):
  - **Only 0-1 constraints**
  - **Summing to 1 (on same layers)**
  - **Euclidean: squares summing to 1 (reminiscent of the wave function in quantum mechanics ~ spectral approach in finding clusters)**

Algorithmic issues II

- **Proceed recursively bottom-to-top**
- **Sum weighted gain/loss events under each of two different scenarios:**
  - **Head Subject has been inherited from parent**
  - **Head Subject has not been inherited from parent**
- **Upon reaching the root, take that with the minimum summary penalty**

Algorithmic issues III

Lifting: Bottom-up recursion under each of two scenarios, (i) HS inherited from parent or (n) not



Application cases

(G) Reconstruction of gene histories over an evolutionary tree (E. Koonin, P. Kellam et al. 2003-2007)

(Aa) representation of research activities of organizations over an ontology of the domain (S. Nascimento et al. 2009 -)

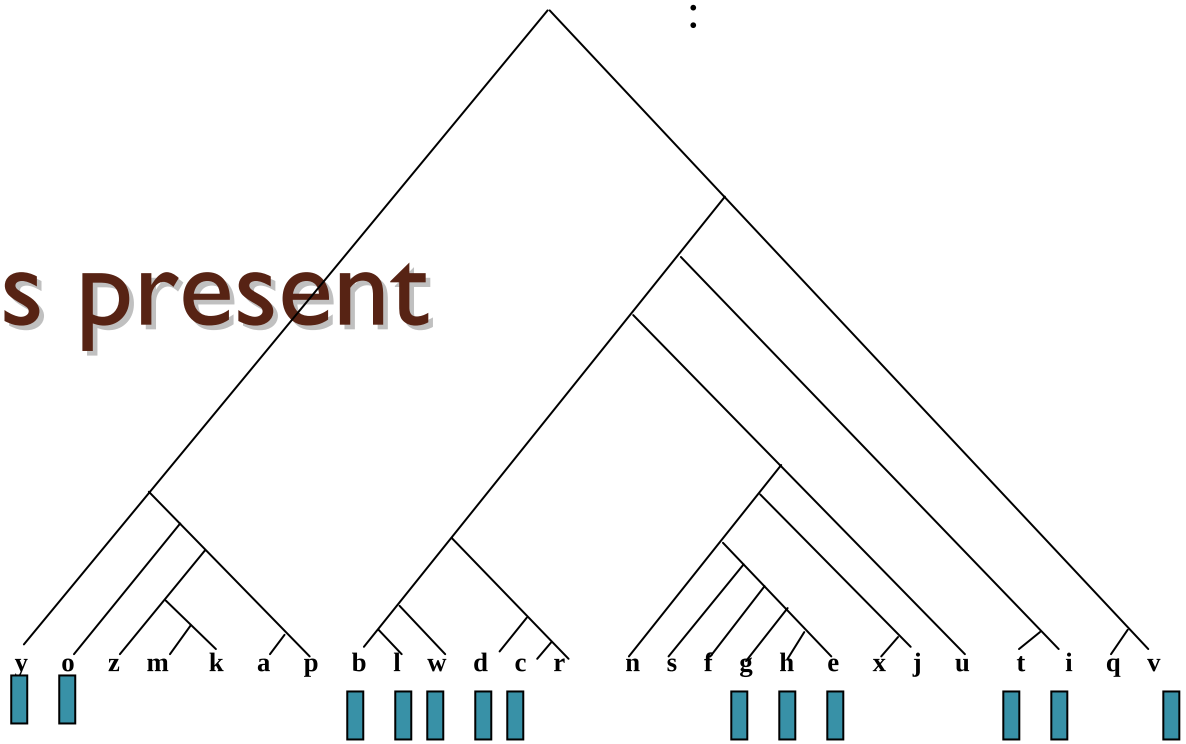
(Ac) Resident complaints management (J. Askarova, E. Babkin, et al., 2011-)

Reconstruction of gene histories over an evolutionary tree

- **Given:**
  - **Evolutionary tree over set of 26 mainly microbial species annotating leaves (at NCBI, USA)**
  - **3166 "COG"s representing individual genes**
- **Problem: Interpret gene histories in tree**
  - **Head subject= Gain of gene, Gap=Loss of gene**
  - **What weights to assign to events?**

Phyletic pattern of COG0572 representing gene "Uridine Kinase" on phylogeny of 26 micro-organisms

- gene is present



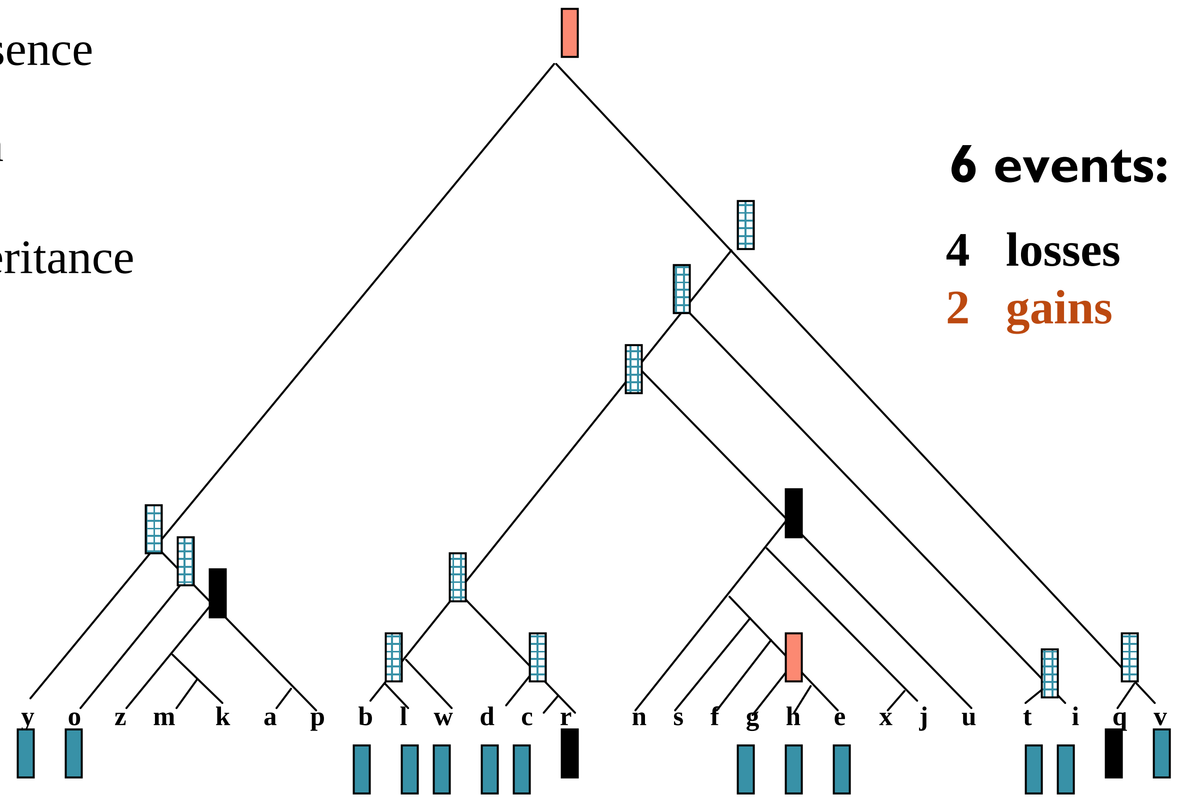
Reconstructed history of COG 0572 Uridine Kinase

presence

gain

inheritance

loss



6 events:

4 losses

2 gains

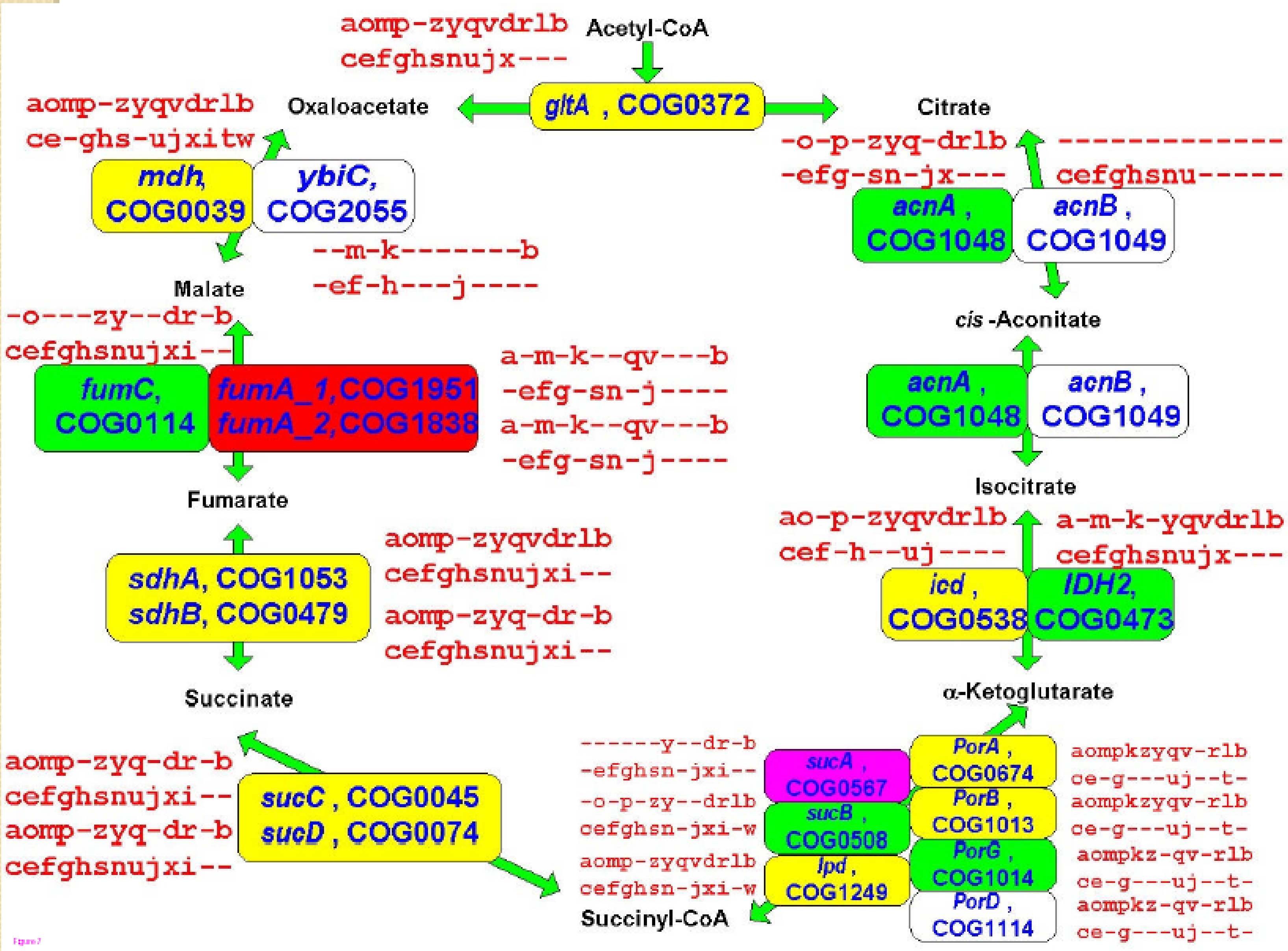
Summary of gene histories at different gain penalties (from 0.1 to 10.0) – which to choose?

At gain penalty 1, 572 gene LUCA is self-sustainable (2003)

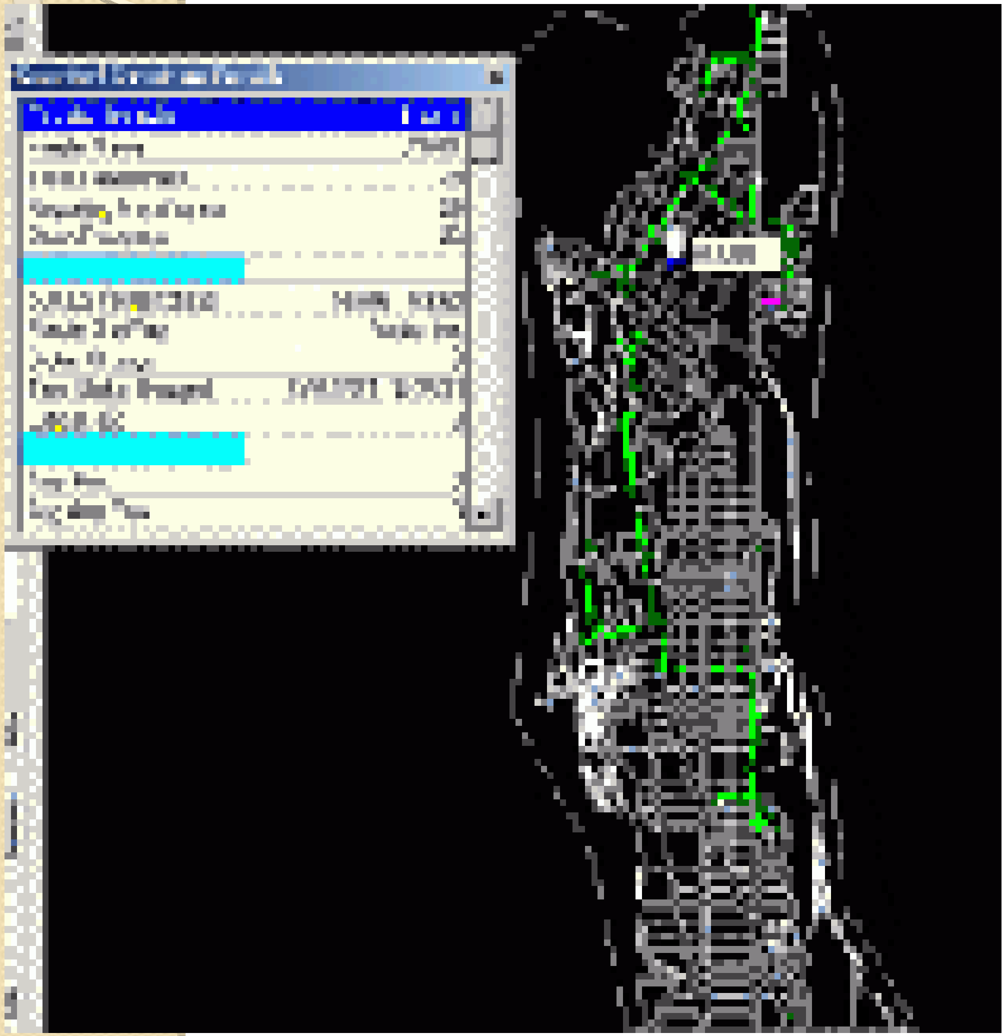
Table 2. Gene sets of ancestral forms and counts of various events in parsimonious scenarios depending on the gain penalty.

Gain penalty (g)	Number of gene (COGs)																
	.1	.2	.3	.4	.5	.6	.7	.8	.9	1	1.25	1.5	2	3	5	7	10
LUCA	84	98	109	132	212	214	266	285	310	572	623	735	956	1211	1325	1664	1725
Archaeal ancestor	390	391	427	521	663	660	732	727	750	977	982	1046	1178	1293	1508	1619	1673
Bacterial ancestor	169	193	243	283	397	415	476	506	532	773	841	986	1259	1582	1879	2028	2091
Horizontal transfer	13241	13001	12315	11464	9462	9312	8733	8363	8315	5495	5136	4238	2646	1293	368	97	13
Loss	0	45	220	512	1506	1595	1989	2259	2301	5121	5562	6872	9944	13695	17535	19230	19947
Total events	16407	16212	15701	15142	14134	14073	13878	13790	13782	13782	13864	14276	15756	18136	21069	22493	23126
Single scenarios	3166	3139	3164	3147	2494	3144	3166	3152	3166	1806	388	2894	2399	2587	2982	3081	3154

Metabolic pathways in LUCA: TCA cycle



(Aa) Representation of activities
Example: for running control



Energy network of Con Edison Company on Manhattan New-York USA (visualized by Advanced Visual Systems company)
to control the energy supply by following all maintenance and repair issues on-line.

Main ingredients:

(i) District **map**,
(ii) Energy network **units**
(iii) **Mapping** (2) at (1).

(Aa) Representation of a Computer Science Department research activities for strategic control

Similar:

- (i') District **Map**: an ontology of Computer Science (CS),
- (ii') Energy maintenance **Units**: clusters of CS research subjects being developed by members of the department,
- (iii') **Mapping** of the research onto the ontology

Member of Department ESSA survey output: Fuzzy membership

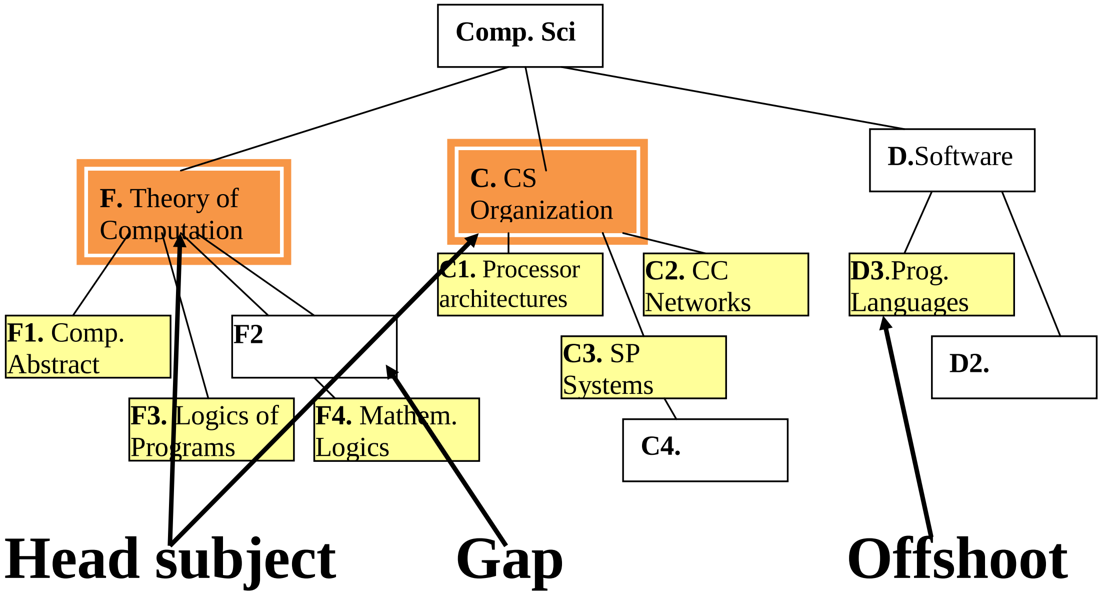
My research topics

Code	Name	%	Options
I.2.6	Learning	30	
I.5.3	Clustering	30	
H.2.8	Database applications	20	
H.5.2	User interfaces	20	
Total weight:		100 %	

Save

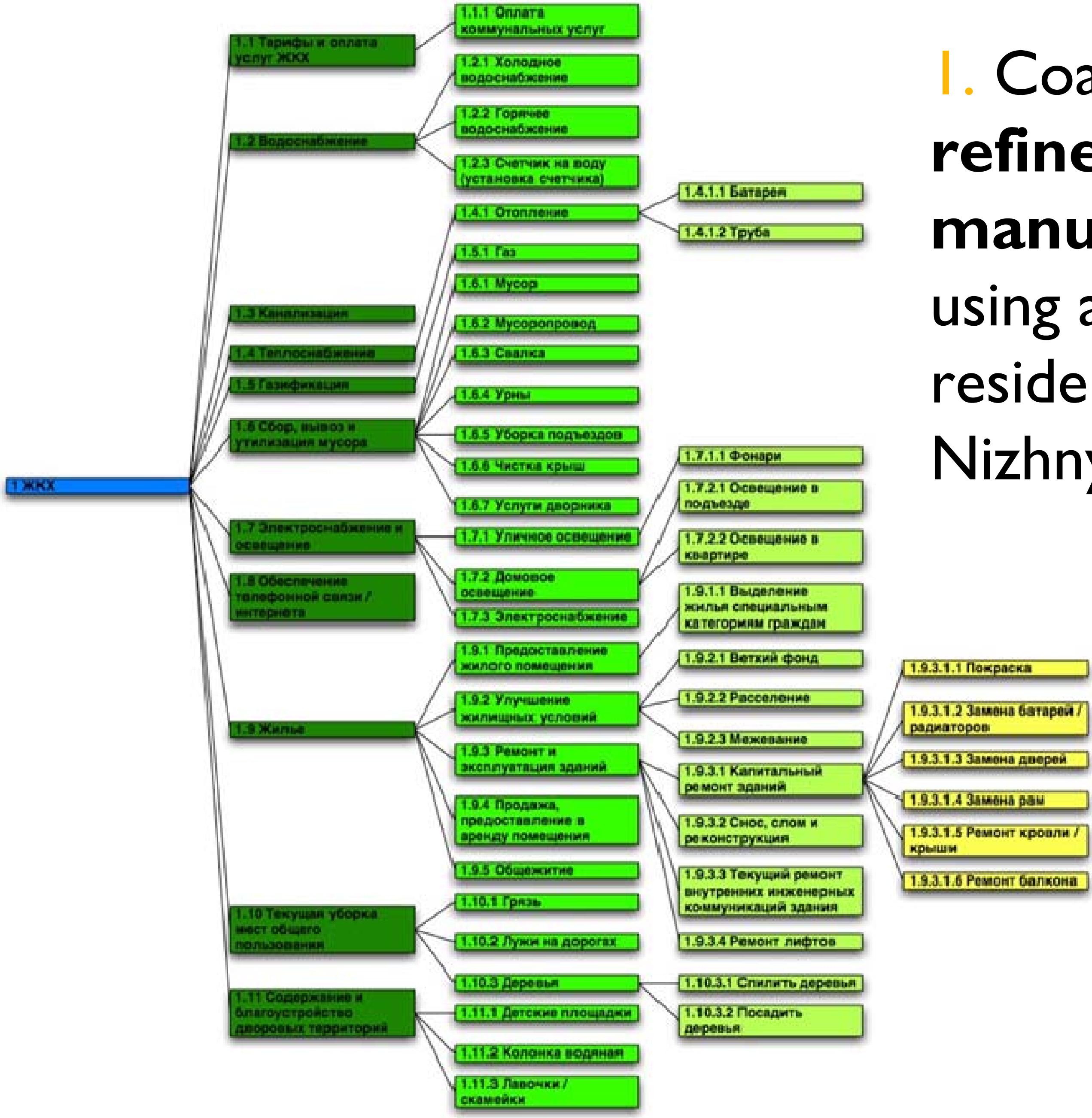
(Ab) An example of annotating a research project

Subject cluster {C1, C2, C3, D3, F1, F3, F4} according to working a team in the department



- Lifting
- Two **Head Subjects**: probably a breaking-through research, say, in distributed logic programming

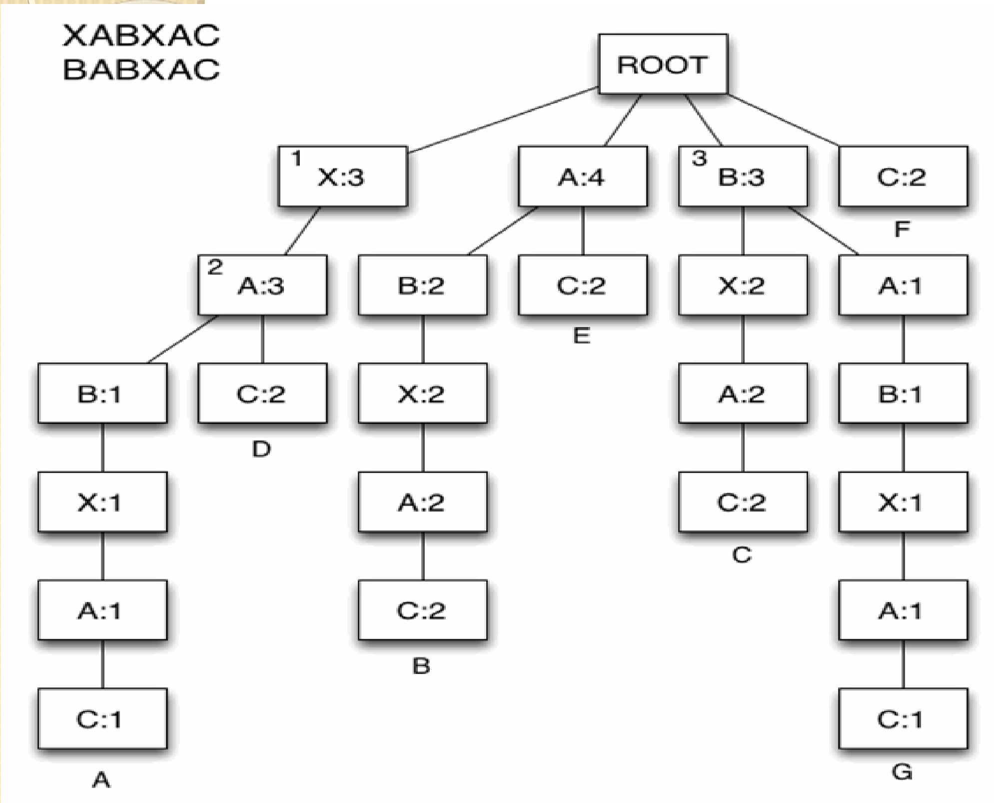
(Ac) Resident complaints management 1



1. Coarse taxonomy **refined, semi-manually** using a database of residents complaints in Nizhny Novgorod

In-house phrase-to-text similarity score:

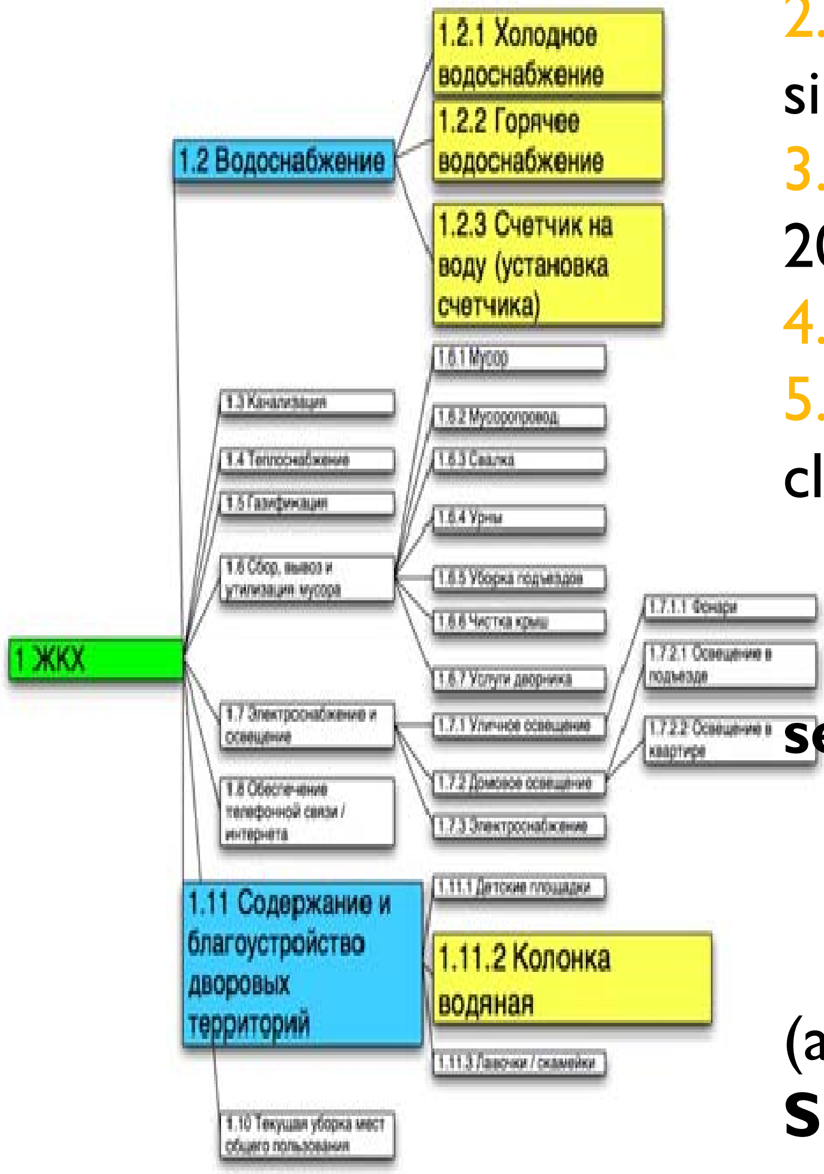
AST symbol's averaged conditional frequency



Suffix tree for strings **XABXAC** and **BABXAC** annotated with substring frequencies, and the **similarity score** for string **VXACA**

Suffix	Match	Score
'VXACA'	**None**	**0**
'XACA'	**'X'->'A'->'C'**	**3/12 + 3/3 + 2/3=1 11/12**
'ACA'	**'A'->'C'**	**4/12 + 2/4=5/6**
'CA'	**'C'**	**2/12**
'A'	**A'**	**4/12**

(Ac) Resident complaints management 2



2. Complaint-to-Topic suffix tree based similarity table S
3. Clusters over S with iK-Means (Mirkin 2012) - Anomalous patterns one-by-one
4. Removal of small and large clusters
5. Parsimoniously lifting remaining clusters

Figure caption:

Cluster mapped to **1. Housing services:**

1.2.1. Hot water problems

1.2.2. Cold water problems

1.2.3. Water meter problems

(all three are parts of **1.2. Water Supply**)

1.11.2. Public water pump

(part of **1.11. Urban landscaping and**

(Ac) Resident complaints management 3

6. Interpretation and conclusions

Observation: Clusters are mapped to overly high ranks

Since the housing and communal services are structured according to technology (water, electricity, public transportation, etc.),

whereas complaints are structured according to living conditions, the latter are frequently at odds with the former:

Organize municipal centers to listen to residents and form multiple-address solutions

(this already is being organized in Moscow, by themselves: no our advice)

Conclusion I

An attempt at a system for computational interpretation – Basic vertical:

- Annotating a single element
- Annotating a granular query set by a single concept
- Annotating a thematic query set within a taxonomy*

*Partly described in

B. Mirkin (2012) Clustering: A Data Recovery Approach, CRC Press.

Conclusion II

Future work

Building taxonomies

“Horizontal” interpretation

Moving to maximum likelihood (via estimation of probabilities using PARL)

Text analysis using more data (string + ”grammar” + net)

Apply to texts, medicine records, documents

Modeling cognitive systems

Similarity between ACMC subjects: example I

ACMC subjects: i, ii, iii, iv, v, vi

Chosen subject memberships for four members

i	.6			.2
ii	.4		.2	.2
iii		.2	.4	.2
iv		.3	.4	.2
v		.5		.2
vi				
	2/5	3/5	3/5	5/5 – member weights

weight = number_of_subjects / max_number_of_subjects

Similarity between ACMC subjects: example 2

i .36 .24 0 0 0 0 0 0 0 0

ii .24 .16 0 0 0 0 0 0 0 0

iii 0.4* 0 0 0 0 0 + 0.6* 0 0 .04 .06 .10 +
0.6* .... =

iv 0 0 0 0 0 0 0 .06 .09 .15

v 0 0 0 0 0 0 0 .10 .15 .25

1st member's 2d member's

i	**0.184**	**0.136**	0.040	0.040	0.040
ii	**0.136**	**0.128**	**0.088**	**0.088**	0.040
iii	0.040	**0.088**	**0.160**	**0.172**	**0.100**
iv	0.040	**0.088**	**0.172**	**0.190**	**0.130**
v	0.040	0.040	**0.100**	**0.130**	**0.190**

not_diagonal_mean =0.0874

Additive fuzzy clustering

Observed:

- **Similarity $B=(b_{ij})$, $i,j\in I$**

To be found:

- **Cluster membership $u=(u_i)$**
- **Intensity $\mu > 0 \Rightarrow$**

Fuzzy cluster similarity $A= \mu^2 uu^T$

K clusters:

$B = A_g + A_1 + A_2 + \ldots + A_K + E$ (g- universal background)

$<B'- A_k, B'- A_k> \Rightarrow \min_{\mu,u}$ one-by-one

Additive fuzzy clustering

- Model: Similarity B summarizes:
  - Background cluster g (all entities)
  - K fuzzy clusters (K unknown)
  - residuals E

$$\mathbf{B = A_g + A_1 + A_2 + \ldots + A_K + E}$$

E to be least-squares minimized over unknown clusters

Method: One cluster at a time

- **Min**$_{\mathbf{u},\,\xi}$ $\sum_{t,t'\in T} (w_{tt'} - \xi u_t u_{t'})2$ $\Leftrightarrow$

- Equivalent to Rayleigh quotient

Max $\mathbf{uWu^T/(u^Tu)}$

- **Spectral approach:** find max eigenvalue and its vector, adjust the latter to fuzzy membership