rocha@lanl.gov

Database Technology for Bioinformatics

From Information Retrieval to Knowledge Systems

Luis M. Rocha
Complex Systems Modeling
CCS3 - Modeling, Algorithms, and Informatics
Los Alamos National Laboratory, MS B256
Los Alamos, NM 87545
rocha@lanl.gov or rocha@santafe.edu

rocha@lanl.gov

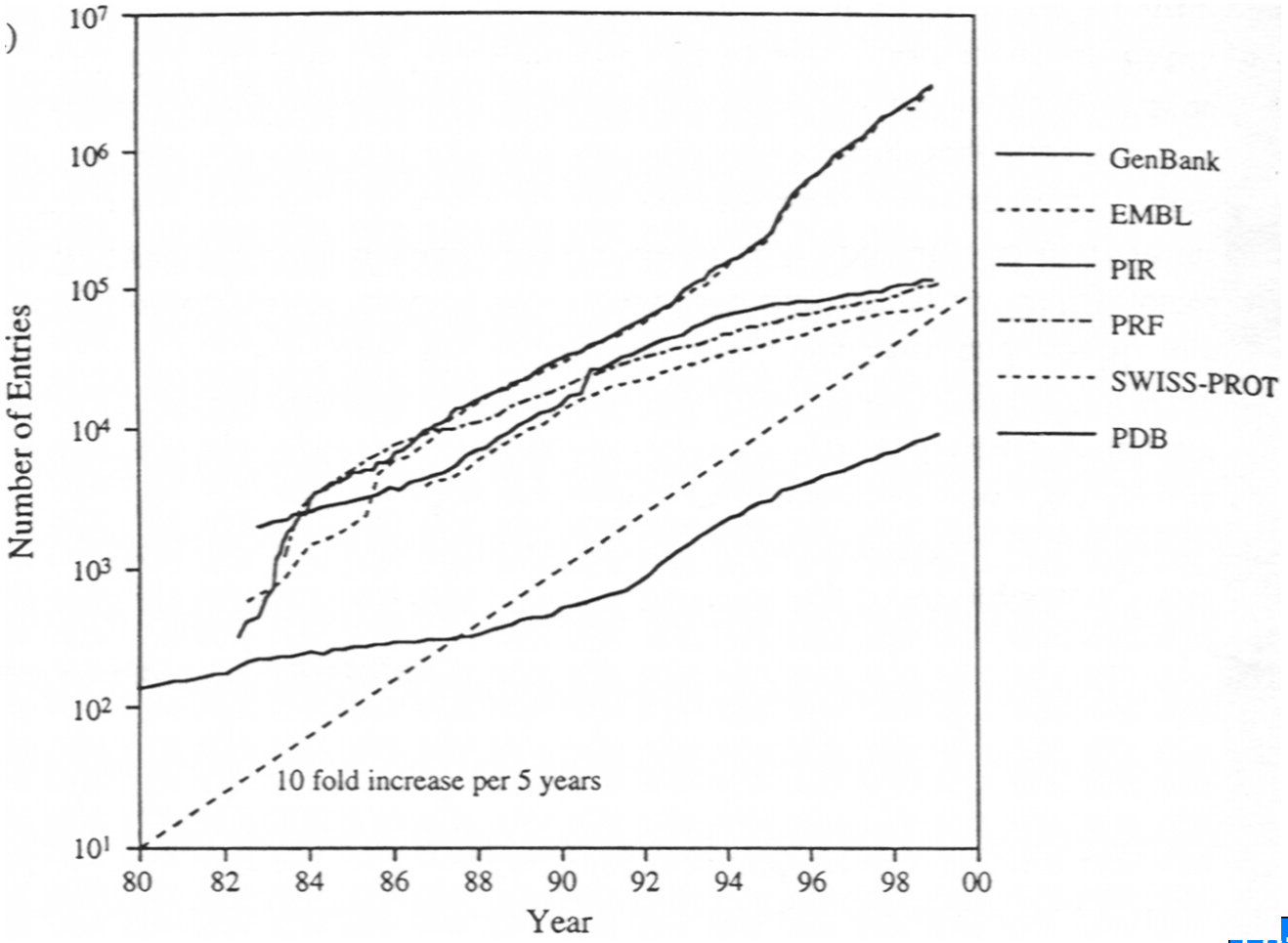
Molecular Biology Databases

- Bibliographic databases
  - On-line journals and bibliographic citations
    - MEDLINE (1971, www.nlm.nih.gov)
- Factual databases
  - Repositories of Experimental data associated with published articles and that can be used for computerized analysis
    - Nucleic acid sequences: GenBank (1982, www.ncbi.nlm.nih.gov), EMBL (1982, www.ebi.ac.uk), DDBJ (1984, www.ddbj.nig.ac.jp)
    - Amino acid sequences: PIR (1968, www-nbrf.georgetown.edu), PRF (1979, www.prf.op.jp), SWISS-PROT (1986, www.expasy.ch)
    - 3D molecular structure: PDB (1971, www.rcsb.org), CSD (1965, www.ccdc.cam.ac.uk)
  - Lack standardization of data contents
- Knowledge Bases
  - Intended for automatic inference rather than simple retrieval
    - Motif libraries: PROSITE (1988, www.expasy.ch/sprot/prosite.html)
    - Molecular Classifications: SCOP (1994, www.mrc-lmb.cam.ac.uk)
    - Biochemical Pathways: KEGG (1995, www.genome.ad.jp/kegg)
  - Difference between knowledge and data (semiosis and syntax)??

rocha@lanl.gov

Growth of sequence and 3D Structure databases

Number of Entries



http://www.c3.lanl.gov/~rocha[/ps/Gordon01.pdf]

Luis Rocha
2001

rocha@lanl.gov

Database Technology and Bioinformatics

- Databases
  - Computerized collection of data for Information Retrieval
  - Shared by many users
  - Stored records are organized with a predefined set of data items (attributes)
  - Managed by a computer program: *the database management system*
  - Schema is the specification of a logical structure of the relations among records and attributes
- Role of databases in Bioinformatics
  - From the dissemination of published work to assisting on-going technology, and, more recently, collaborative research
  - Essential aspect of Bioinformatics needed to manage large-scale projects and heterogeneous research groups
- Flat File Databases
  - Sequential collection of entries, stored in a set of text files
  - Easy for programs to handle and utilize data

rocha@lanl.gov

Example of Flat File Sequence Database

GenBank (a)

```
LOCUS       DRODPPC      4001 bp    mRNA            INV       15-MAR-1990
DEFINITION  D.melanogaster decapentaplegic gene complex (DPP-C), complete cds.
ACCESSION   M30116
NID         g157291
KEYWORDS    .
SOURCE      D.melanogaster, cDNA to mRNA.
  ORGANISM  Drosophila melanogaster
            Eukaryotae; mitochondrial eukaryotes; Metazoa; Arthropoda;
            Tracheata; Insecta; Pterygota; Diptera; Brachycera; Muscomorpha;
            Ephydroidea; Drosophilidae; Drosophila.
REFERENCE   1  (bases 1 to 4001)
  AUTHORS   Padgett,R.W., St Johnston,R.D. and Gelbart,W.M.
  TITLE     A transcript from a Drosophila pattern gene predicts a protein
            homologous to the transforming growth factor-beta family
  JOURNAL   Nature 325, 81-84 (1987)
  MEDLINE   87090408
COMMENT     The initiation codon could be at either 1188-1190 or 1587-1589.
```

rocha@lanl.gov

Example of Flat File Sequence Database

GenBank (b)

```
FEATURES             Location/Qualifiers
     source          1..4001
                     /organism="Drosophila melanogaster"
                     /db_xref="taxon:7227"
     mRNA            <1..3918
                     /gene="dpp"
                     /note="decapentaplegic protein mRNA"
                     /db_xref="FlyBase:FBgn0000490"
     gene            1..4001
                     /note="decapentaplegic"
                     /gene="dpp"
                     /allele=""
                     /db_xref="FlyBase:FBgn0000490"
     CDS             1188..2954
                     /gene="dpp"
                     /note="decapentaplegic protein (1188 could be 1587)"
                     /codon_start=1
                     /db_xref="FlyBase:FBgn0000490"
                     /db_xref="PID:g157292"
                     /translation="MRAWLLLLAVLATFQTIVRVASTEDISQRFIAAIAPVAAHIPL
                     SASGSGSGRSGSRSVGASTSTALAKAFNPFSEPASFSDSDKSHRSKTNKKPSKSDAN
                     ..........
                     LGYDAYYCHGKCPFPLADHFNSTNHAVVQTLVNNMNPGKVPKACCVPTQLDSVAMLY
                     NDQSTVVLKNYQEMTVVGCGCR"
```

Luis Rocha
2001

http://www.c3.lanl.gov/~rocha[/ps/Gordon01.pdf]

Los Alamos
National Laboratory

rocha@lanl.gov

Example of Flat File Sequence Database

GenBank (a)

```
BASE COUNT     1170 a    1078 c     956 g     797 t
ORIGIN
        1 gtcgttcaac agcgctgatc gagtttaaat ctataccgaa atgagcggcg gaaagtgagc
       61 cacttggcgt gaacccaaag ctttcgagga aaattctcgg acccccatat acaaatatcg
      121 gaaaaagtat cgaacagttt cgcgacgcga agcgttaaga tcgccaaaag atctccgtgc
      181 ggaaacaaag aaattgaggc actattaaga gattgttgtt gtgcgcgagt gtgtgtcttc
      241 agctgggtgt gtggaatgtc aactgacggg ttgtaaaggg aaaccctgaa atccgaacgg
      301 ccagccaaag caaataaagc tgtgaatacg aattaagtac aacaaacagt tactgaaaca
      361 gatacagatt cggattcgaa tagagaaaca gatactggag atgccccag aaacaattca
      421 attgcaaata tagtgcgttg cgcgagtgcc agtggaaaaa tatgtggatt acctgcgaac
      481 cgtccgccca aggagccgcc gggtgacagg tgtatccccc aggataccaa cccgagccca
      541 gaccgagatc cacatccaga tcccgaccgc agggtgccag tgtgtcatgt gccgcggcat
      601 accgaccgca gccacatcta ccgaccaggt gcgcctcgaa tgcggcaaca caattttcaa
          ..........
     3841 aactgtataa acaaaacgta tgccctataa atatatgaat aactatctac atcgttatgc
     3901 gttctaagct aagctcgaat aaatccgtac acgttaatta atctagaatc gtaagaccta
     3961 acgcgtaagc tcagcatgtt ggataaatta atagaaacga g
//
```

Luis Rocha
2001

http://www.c3.lanl.gov/~rocha[/ps/Gordon01.pdf]

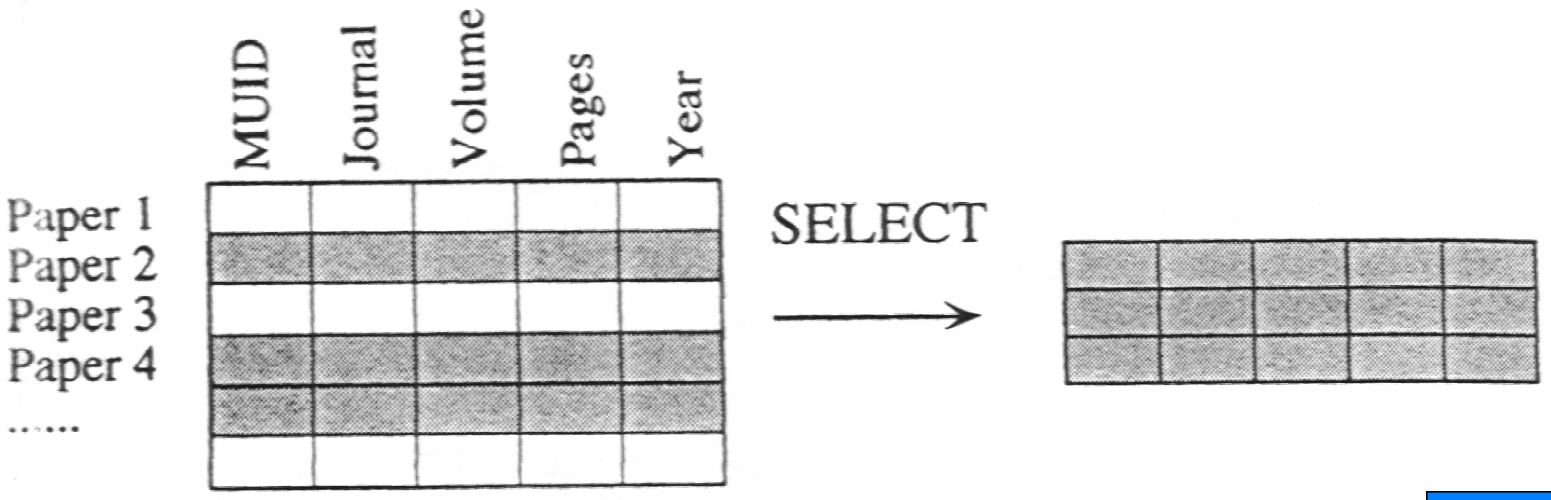
Los Alamos
National Laboratory

rocha@lanl.gov

Relational Databases: Codd, 1970

Relational database management system (RDBMS): Most popular commercial database type.

- All data is organized as tables or relations between records and attributes
  - A relation associates the elements of 2 or more sets
  - Example: records are papers and attributes are bibliographic info
- Selection
  - Selects records according to attribute criteria (row extraction)
  - E.g. papers published in YEAR=x



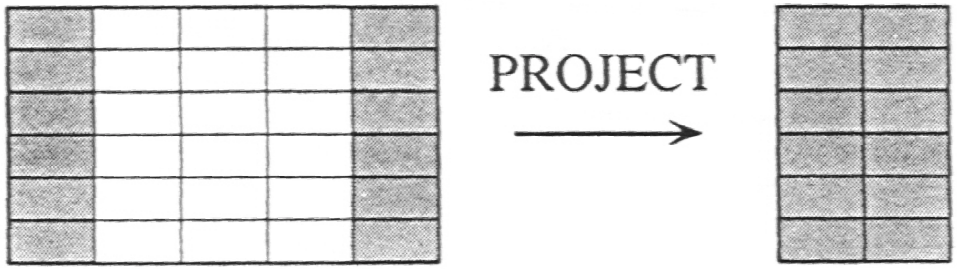
http://www.c3.lanl.gov/~rocha[/ps/Gordon01.pdf]

Luis Rocha
2001

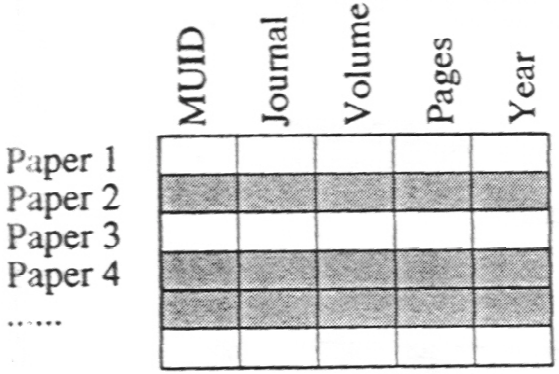
rocha@lanl.gov

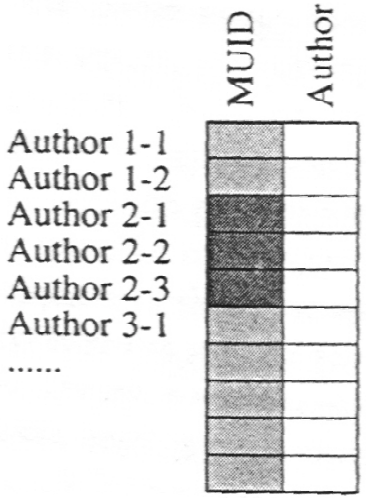
Relational Database

- Projection
  - Extracts columns: Projects a set of papers into a reduced set of attributes.



- Relational databases contain several tables (multiple relations)





http://www.c3.lanl.gov/~rocha[/ps/Gordon01.pdf]

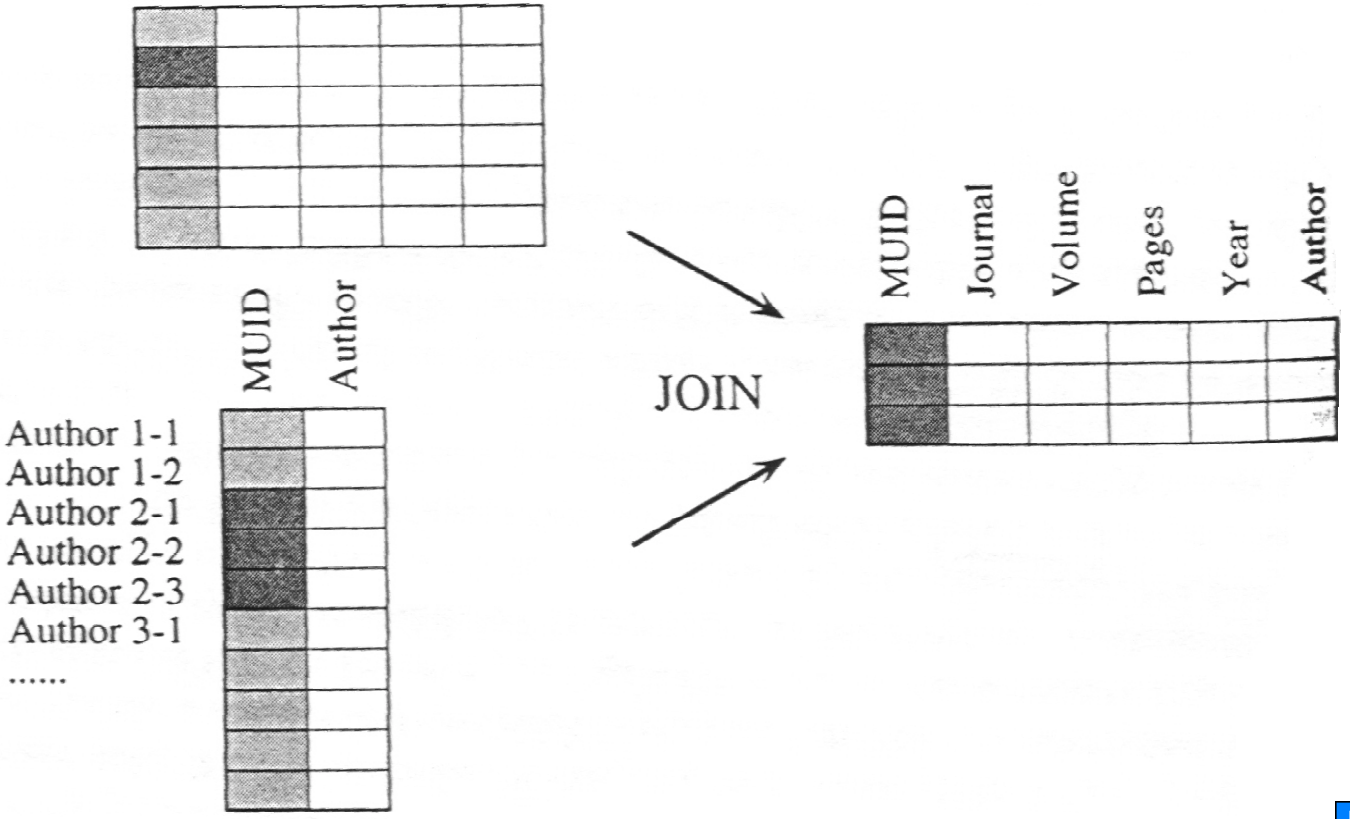
Luis Rocha
2001

Los Alamos
National Laboratory

rocha@lanl.gov

Relational database

- Join
  - Merges records that contain matching values for specified attributes
  - Example: given a value of ID join records from both tables



http://www.c3.lanl.gov/~rocha[/ps/Gordon01.pdf]

Luis Rocha
2001

rocha@lanl.gov

Relational Algebra

Set of Operations used for IR

- Select, Project, Join
  - Implemented in some language such as SQL (Structured Query Language)
    - Example: SELECT * FROM CITATION_TABLE WHERE PUBLISHED_YEAR='1995'
    - * Denotes ALL.

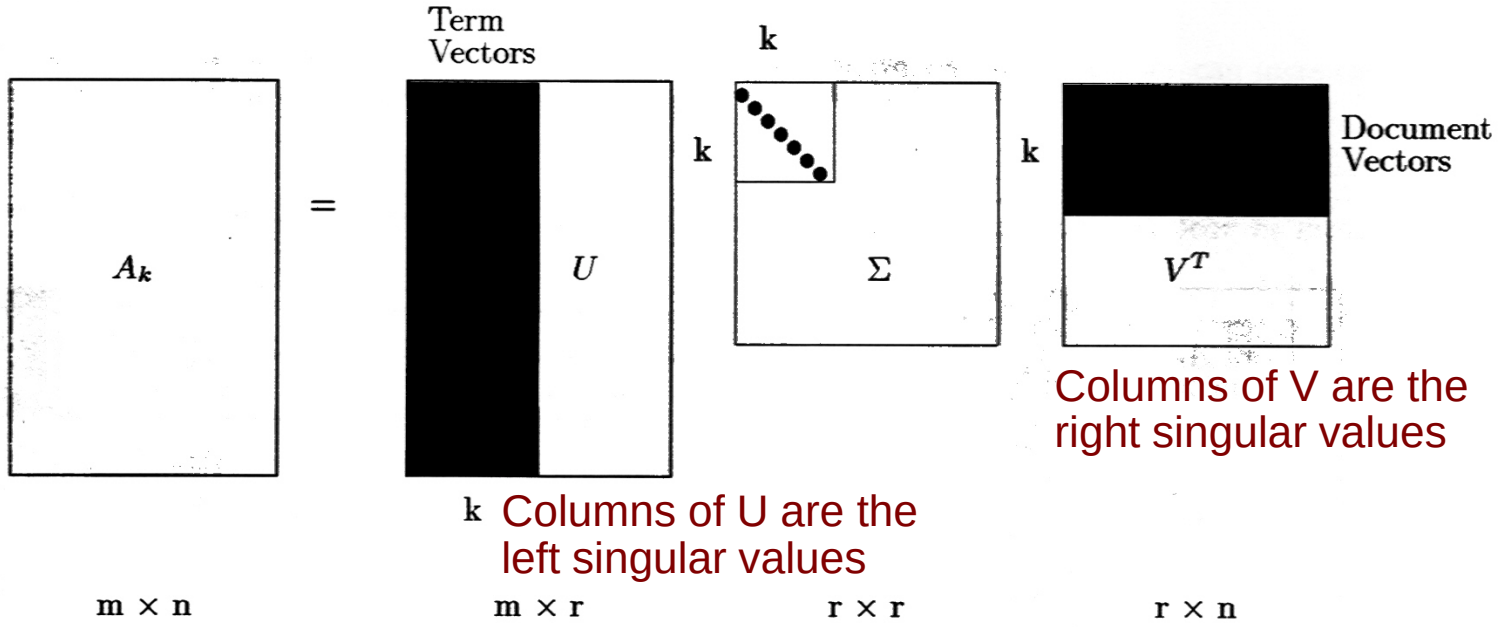
But Relational databases are not very good at representing hierarchical aspects of the real world. It also typically requires many tables to represent a data enevironment.

- Object-oriented databases are typically more flexible, but there are not many standard comercial products available
  - Hierarchical strcuture of records with associated attributes and code. Inheritable.

rocha@lanl.gov

Latent Databases

Using SVD



SVD allows us to obtain the lower rank approximations that best approximate the original matrix. What is lost by losing weaker singular values, is unnecessary noise. The underlying, essential structure of associations between keyterms and records is kept

Luis Rocha
2001

http://www.c3.lanl.gov/~rocha[/ps/Gordon01.pdf]

Los Alamos
National Laboratory

rocha@lanl.gov

Latent Databases

Keyword × Documents Relation is stored as a lower k SVD representation

Example: Small database from 17 books reviewed by SIAM Review

Underlied words are keyterms

Label	Titles
B1	A Course on Integral Equations
B2	Attractors for Semigroups and Evolution Equations
B3	Automatic Differentiation of Algorithms: Theory, Implementation, and Application
B4	Geometrical Aspects of Partial Differential Equations
B5	Ideals, Varieties, and Algorithms – An Introduction to Computational Algebraic Geometry and Commutative Algebra
B6	Introduction to Hamiltonian Dynamical Systems and the N-Body Problem
B7	Knapsack Problems: Algorithms and Computer Implementations
B8	Methods of Solving Singular Systems of Ordinary Differential Equations
B9	Nonlinear Systems
B10	Ordinary Differential Equations
B11	Oscillation Theory for Neutral Differential Equations with Delay
B12	Oscillation Theory of Delay Differential Equations
B13	Pseudodifferential Operators and Nonlinear Partial Differential Equations
B14	Sinc Methods for Quadrature and Differential Equations
B15	Stability of Stochastic Differential Equations with Respect to Semi-Martingales
B16	The Boundary Integral Approach to Static and Dynamic Contact Problems
B17	The Double Mellin-Barnes Type Integrals and Their Applications to Convolution Theory

Luis Rocha
2001

http://www.c3.lanl.gov/~rocha[/ps/Gordon01.pdf]

rocha@lanl.gov

16×17 Keyterm × Document Matrix

Terms	Documents																
	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12	B13	B14	B15	B16	B17
algorithms	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0
application	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
delay	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
differential	0	0	0	1	0	0	0	1	0	1	1	1	1	1	1	0	0
equations	1	1	0	1	0	0	0	1	0	1	1	1	1	1	1	0	0
implementation	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0
integral	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
introduction	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
methods	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0
nonlinear	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0
ordinary	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0
oscillation	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
partial	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0
problem	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0
systems	0	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0	0
theory	0	0	1	0	0	0	0	0	0	0	1	1	0	0	0	0	1

Think of a database of GEA data sets (instead of or in addition to documents), indexed by relevant genes (instead or in addition to keyterms)

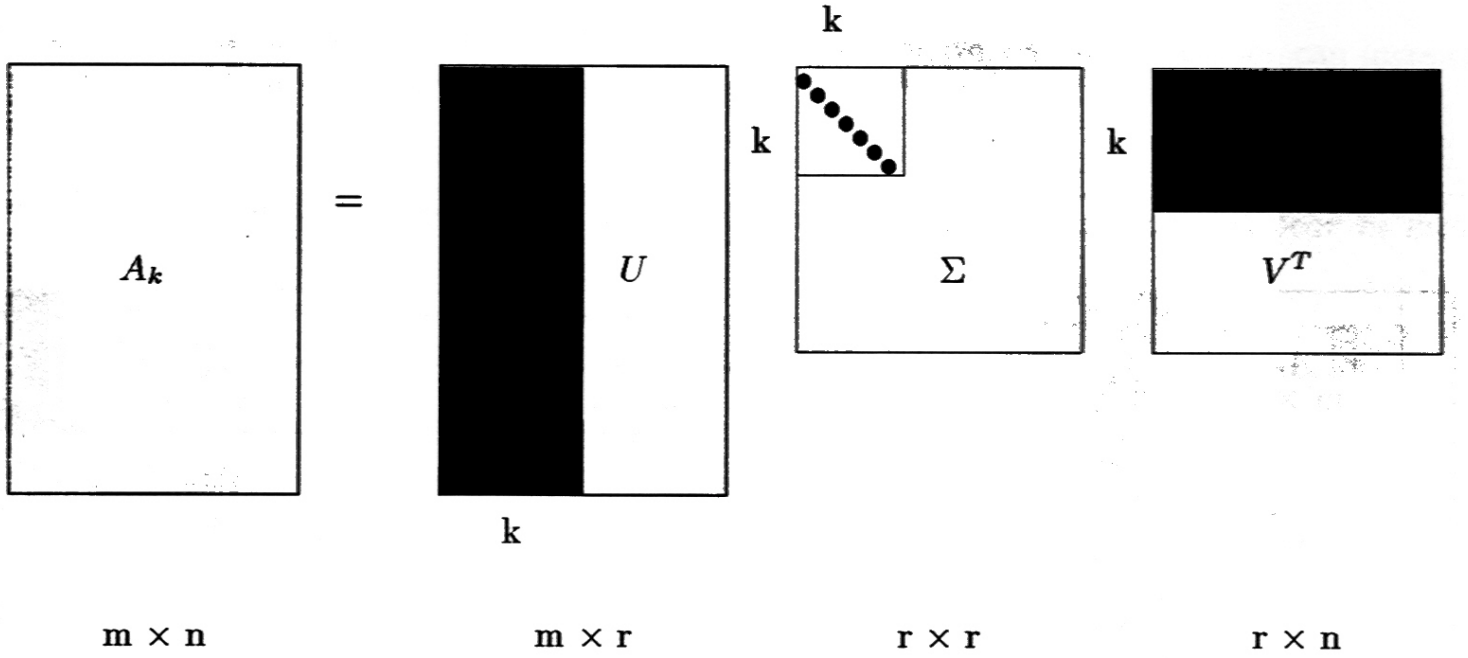
Luis Rocha
2001

http://www.c3.lanl.gov/~rocha[/ps/Gordon01.pdf]

rocha@lanl.gov

Term × Document SVD

m=17, n=16



Columns are terms and rows are documents

Columns of U are eigenterms (rows are documents)

Rows of V^T are eigendocuments (columns are terms)

Luis Rocha
2001

http://www.c3.lanl.gov/~rocha[/ps/Gordon01.pdf]

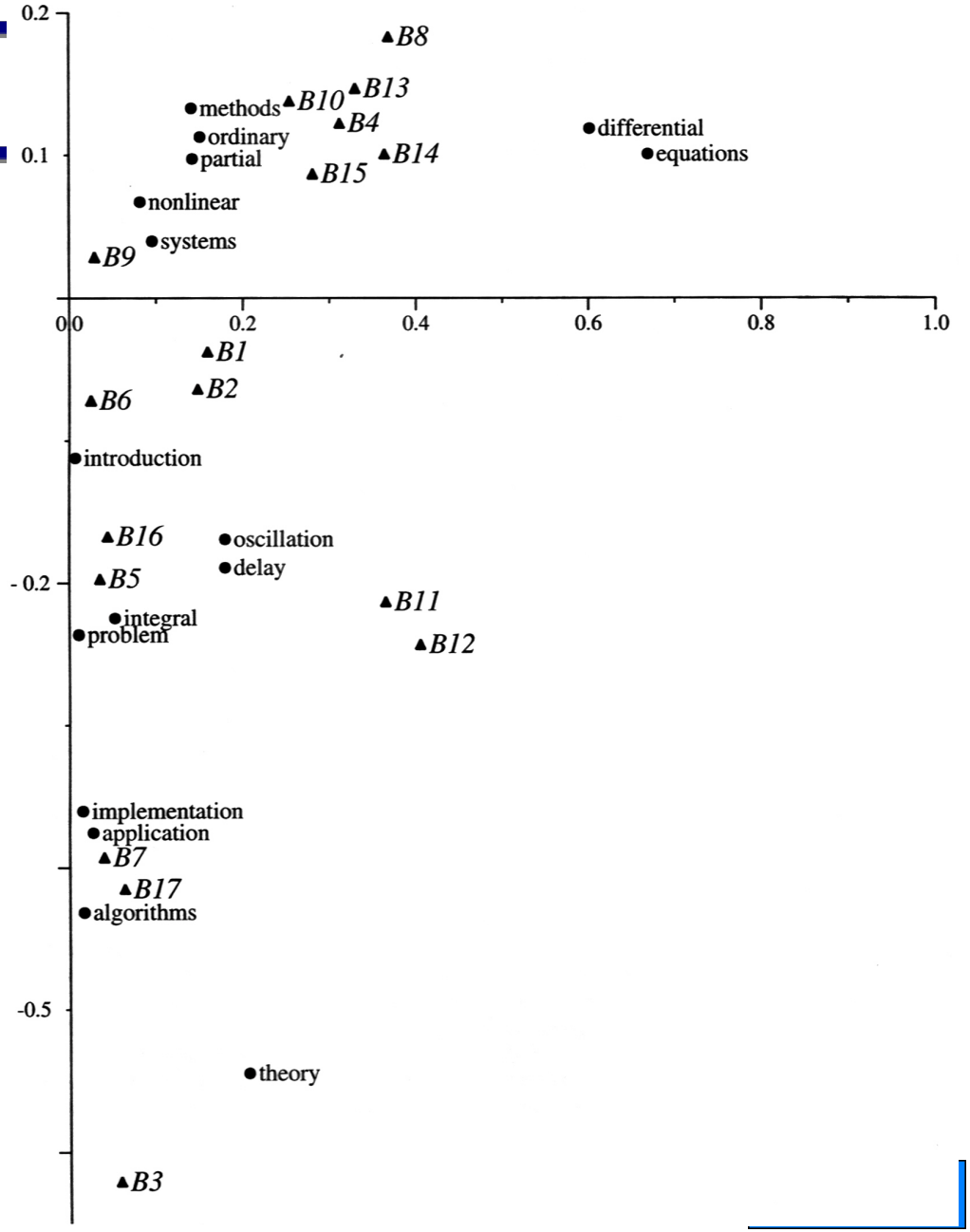
rocha@lanl.gov

SVD Aprox for k=2

Document and terms are plotted according to coefficients in the derived 2 eigenterms and eigendocuments

X seems to be about "differential equations" while y about more general algorithms and applications

Again think of the gene/dataset analogy



Luis Rocha
2001

http://www.c3.lanl.gov/~rocha[/ps/Gordon01.pdf]

rocha@lanl.gov

Keyterm Query

$$q_k = q^T V_k \Sigma_k^{-1}$$

Eigendocuments

$$\begin{pmatrix} 0.0511 & -0.3337 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}^T \begin{pmatrix} 0.0159 & -0.4317 \\ 0.0266 & -0.3756 \\ 0.1785 & -0.1692 \\ 0.6014 & 0.1187 \\ 0.6691 & 0.1209 \\ 0.0148 & -0.3603 \\ 0.0520 & -0.2248 \\ 0.0066 & -0.1120 \\ 0.1503 & 0.1127 \\ 0.0813 & 0.0672 \\ 0.1503 & 0.1127 \\ 0.1785 & -0.1692 \\ 0.1415 & 0.0974 \\ 0.0105 & -0.2363 \\ 0.0952 & 0.0399 \\ 0.2051 & -0.5448 \end{pmatrix} \begin{pmatrix} 4.5314 & 0 \\ 0 & 2.7582 \end{pmatrix}^{-1}$$

Terms

Query coordinates in reduced space

singular values

Luis Rocha
2001

http://www.c3.lanl.gov/~rocha[/ps/Gordon01.pdf]

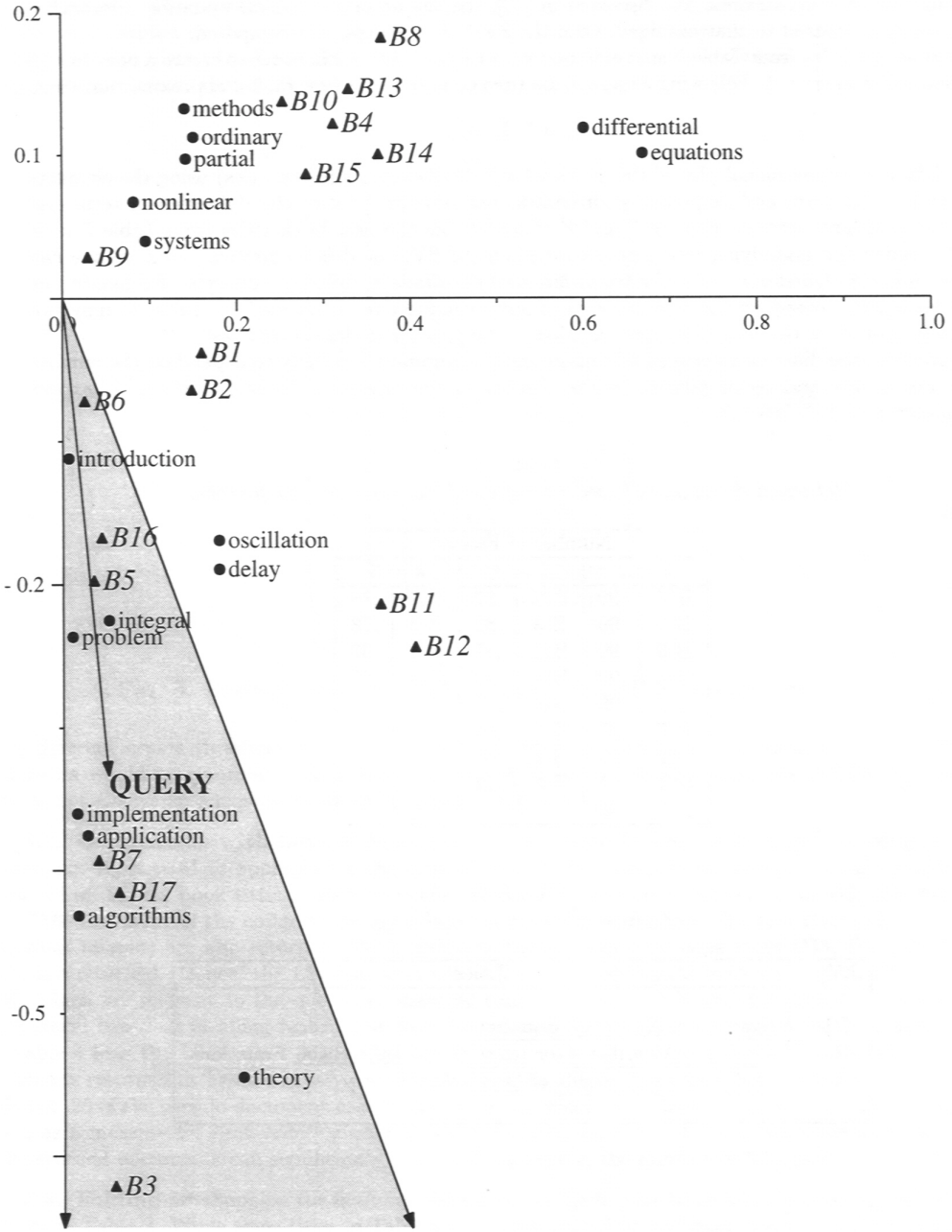
rocha@lanl.gov

Keyterm Query

Retrieval

Can also retrieve documents close to a set of other documents

For a database of datasets, this would mean that we would retrieve those data sets most relevant to study the genes in a query



Luis Rocha
2001

http://www.c3.lanl.gov/~rocha[/ps/Gordon01.pdf]

rocha@lanl.gov

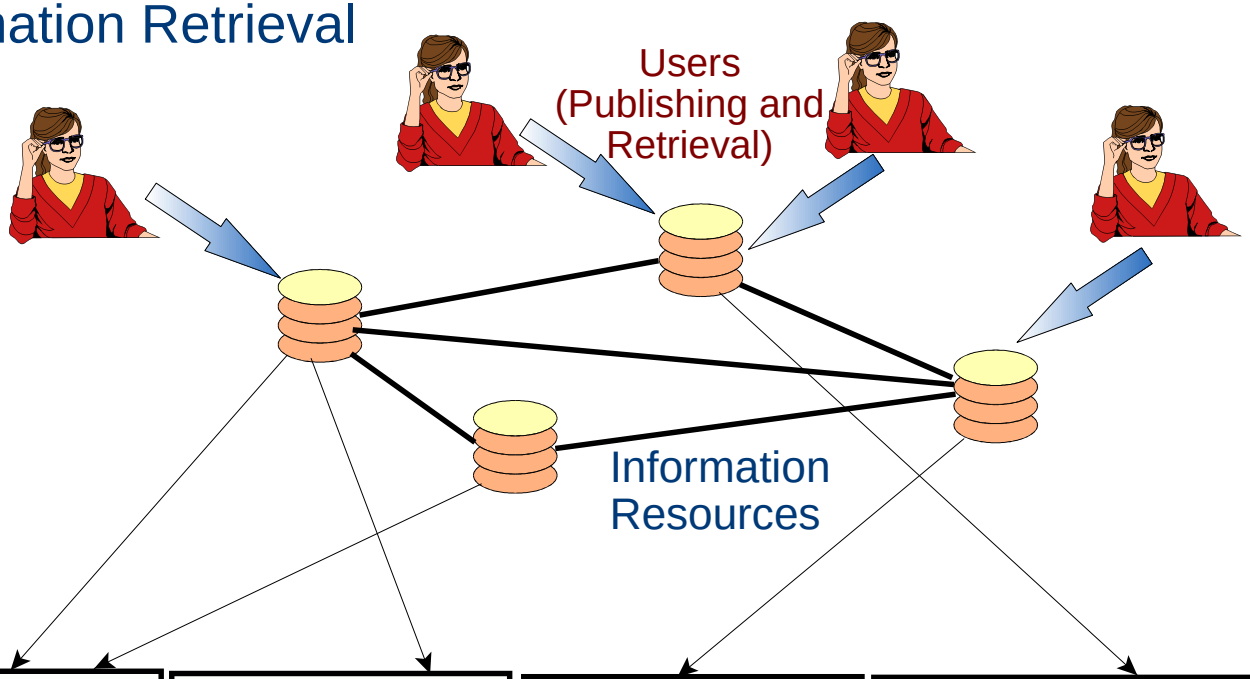
References

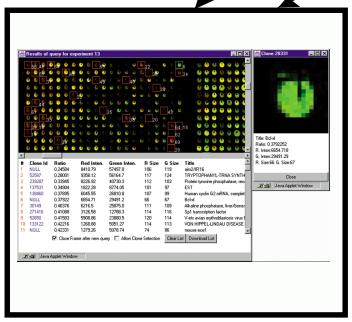
- SVD and Latent Semantic Analysis in IR
  - Berry, M.W., S.T. Dumais, and G.W. O'Brien [1995]."Using linear algebra for intelligent information retrieval." *SIAM Review*. Vol. 37, no. 4, pp. 573-595.
  - Kannan, R. and S. Vempala [1999]."Real-time clustering and ranking of documents on the web." Unpublished Manuscript.
  - Landauer, T.K., P.W. Foltz, and D. Laham [1998]."Introduction to Latent Semantic Analysis." *Discourse Processes*. Vol. 25, pp. 259-284.
- SVD for Gene Expression Analysis
  - Alter, O., P.O. Brown and D. Botstein [2000]."Singular value decomposition for genome-wide expression data processing and modeling." *PNAS*. Vol. 97, no. 18, pp. 10101-06.
  - Hastie, T. et al [2000]."'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns." *Genome Biology*. Vol. 1, no. 2, pp. 3.1-3.21.
  - Holter, N.S. et al [2000]."Fundamental patterns underlying gene expression profiles:Simplicity from complexity." *PNAS*. Vol. 97, no. 15, pp. 8409-14.
  - Raychaudhuri, S., J.M. Stuart and R.B. Altman [2000]."Principal components analysis to summarize microarray experiments: Application to sporulation data." ????? http://cmgm.stanford.edu.
  - Wall, M., P.A. Dyck, and T. Brettin [2001]."SVDMAN -- Singular value decomposition analysis of microarray data." *Bioinformatics*. In Press.

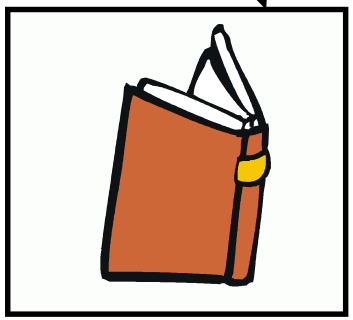
rocha@lanl.gov

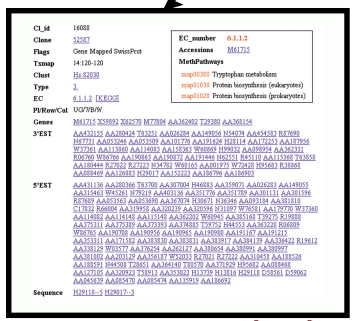
Scientific Databases

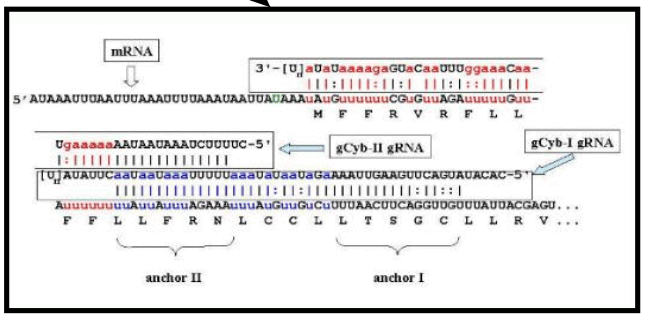
Information Retrieval











Datasets, analysis results, hypothesis, publications, algorithms, simulations, software, ideas

http://www.c3.lanl.gov/~rocha[/ps/Gordon01.pdf]

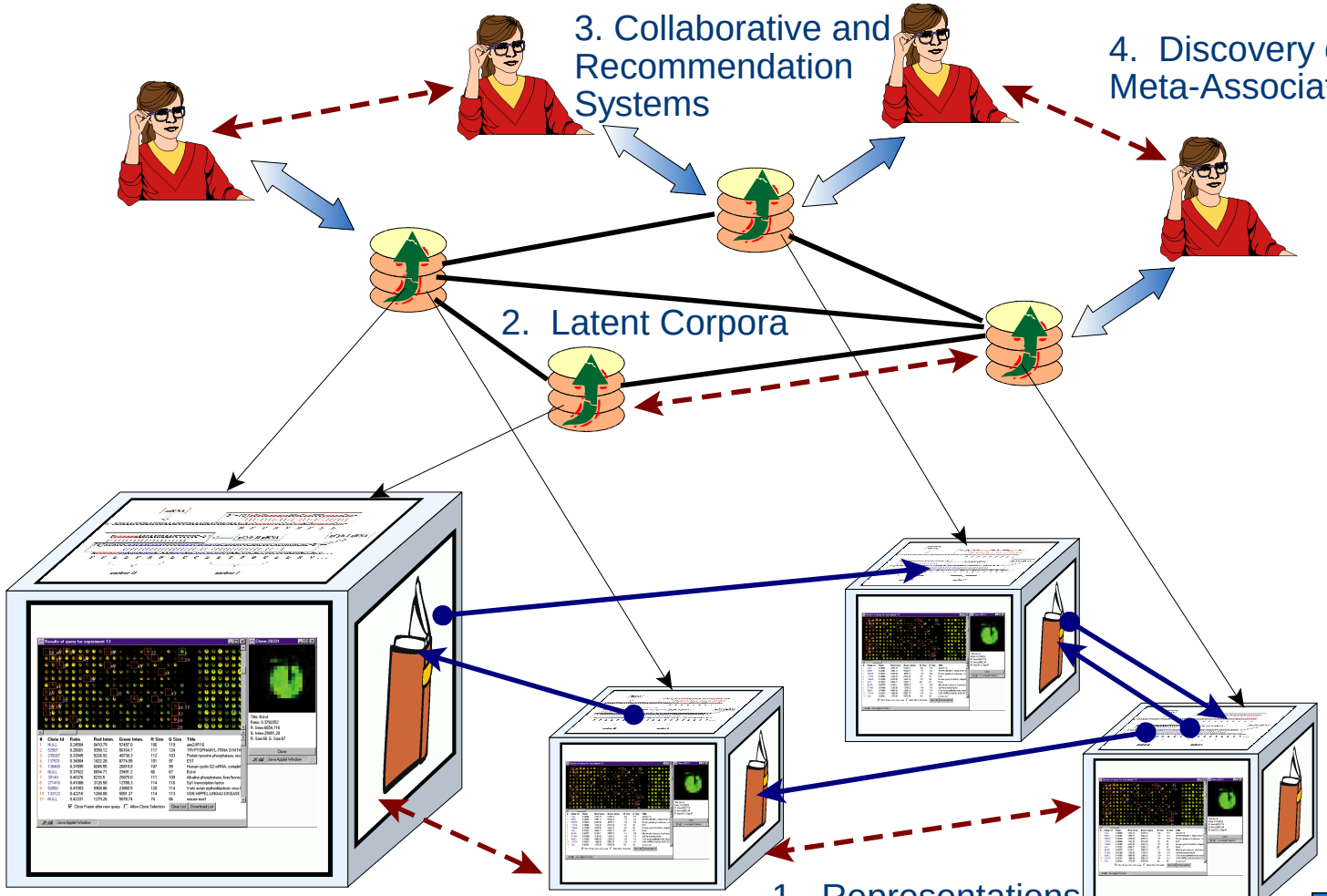
Luis Rocha
2001

Los Alamos National Laboratory

rocha@lanl.gov

Distributed Knowledge Systems

Collaborative Scientific Environments



http://www.c3.lanl.gov/~rocha[/ps/Gordon01.pdf]

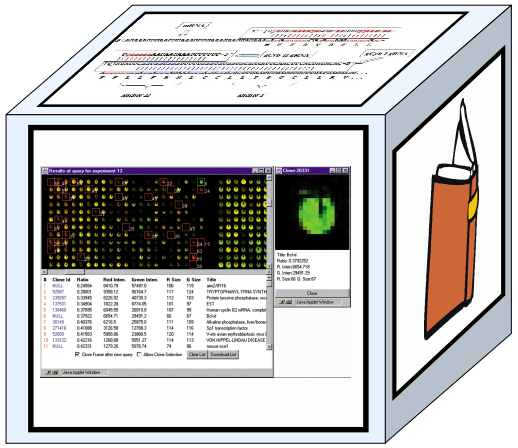
Luis Rocha
2001

Los Alamos National Laboratory

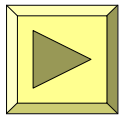
rocha@lanl.gov

1. Representations of Biological Data

Data Objects and Object Architectures



- Objects capable of grouping data sets, reports, code, etc.
  - Networked, proactive containers
  - Nelson's Buckets: Intelligent Data Agents
  - Object Management Group
- Specify specific needs of biological data
- Relates strongly to Computational component



Luis Rocha
2001

http://www.c3.lanl.gov/~rocha[/ps/Gordon01.pdf]

Los Alamos National Laboratory

rocha@lanl.gov

Representations Of Biological Data

Semantic Markup and Exchange Protocols

- To facilitate retrieval, linking, and intelligent behavior of data objects there is a need to characterize data according to the needs of users.
  - Standards based on XML
    - GEML (Gene Expression Markup Language)
    - GeneX (NCGR)
  - Domains can be conceptualized as ontologies
    - Bio-ontologies Consortium
    - BioPathways Consortium
  - Exchange protocols
    - Based on RDF(S) (Resource Description Format Schema)
    - Ontology Interchange Layer (OIL)
    - For biological data: EcoCyc and TAMBIS.
- Aim is to select and develop appropriate representation for our data.

rocha@lanl.gov

XML Repository

Example of a Record

```
<RECORD>
        <ARPID>ISSN_1013-9826_1998_137_55</ARPID>
        <CITATION>
            <REFID>ISSN_0032-3861_1987_28_1489 <~>
            ISSN_0022-2461_1994_29_3377 <~>ISSN_0032-3861_1994_35_3948 <~>
            ISSN_0032-3861_1995_36_4587 <~> ISSN_0032-3861_1995_36_4605 <~>
            ISSN_0032-3861_1995_36_4621 <~> ISSN_0022-2461_1989_24_298 <~>
            ISSN_0032-3861_1980_21_466 <~> ISSN_0032-3861_1985_26_1855
          </REFID>
        </CITATION>
        <BIBLIO>
            <TITLE>Effect of rubber functionality on mechanical and fracture properties of impact-modified
nylon 6,6/polypropylene blends</TITLE>
            <ENUM TYPE="ENDPAGE">62</ENUM>
        </BIBLIO>
        <KEYTERMS>
            <KEYW TYPE="TITLE">rubber <~> properti <~> nylon <~> mechan <~> impact-modifi <~> function
<~> fractur <~> blend <~> /polypropylen</KEYW>
            <KEYW TYPE="KEYW_AU">PA6,6/PP blends <~> rubber-toughened nylon <~> rubber-toughened
polypropylene <~> mechanical properties <~> fracture toughness <~> J(c) <~> fractography</KEYW>
            <KEYW TYPE="KEYW_ISI">FILLED COMPOSITE-MATERIALS <~> POLYPROPYLENE BLENDS
<~> BLOCK-COPOLYMERS <~> PREDICTIVE MODEL <~> COMPATIBILIZATION <~> POLYAMIDES <~>
MORPHOLOGY <~> CAVITATION <~> PARTICLES</KEYW>
            <KEYW TYPE="AUTHOR">Wong, SC <~> Mai, YW</KEYW>
        </KEYTERMS>
    </RECORD>
```

rocha@lanl.gov

GEML

Example of a Pattern File

```
<?xml version="1.0" standalone="no"?>
<!DOCTYPE project SYSTEM "GEMLPattern.dtd">
<project name="Hsapiens-421205160837" date="07-12-1999 12:43:48" by="jzsmith" company="JZSmith Technologies" >
    <pattern name="Hsapiens-421205160837">
        <reporter  name="XV186450"  systematic_name="XV186450"
                   active_sequence="TCTCACTGGTCAGGGGTCTTCTCCC" start_coord="159">
            <feature number="6878">
               <position  x="0.3333" y="0.508374" units="inches" />
            </feature>
            <gene primary_name="XV186450" systematic_name="XV186450" >
               <accession database="n/a" id="XV186520" />
            </gene>
        </reporter>
        <reporter name="T89593" systematic_name="T89593"
                   active_sequence="TACAGTGTCAGAATTAACTGTAGTC" start_coord="201"  >
            <feature number="6879">
               <position  x="0.340707" y="0.508374" units="inches" />
            </feature>
            <gene primary_name="T89593" systematic_name="T89593" >
                <accession database="n/a" id="T89593" />
            </gene>
        </reporter>
        <!-- Total Number of Reporters: 2 -->
    </pattern>
    <printing date="07-12-1999 12:43:48" printer="IJS 3" type="INKJET"
              pattern_name="Hsapiens- 421205160837" >
        <chip barcode="JZS123456781" />
        <chip barcode="JZS123456782" />
        <chip barcode="JZS123456783" />
        <chip barcode="JZS123456784" />
     </printing
</project>
```

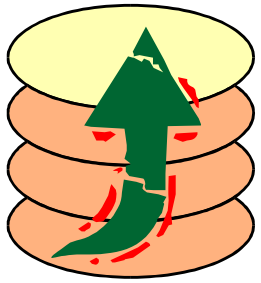

http://www.c3.lanl.gov/~rocha[/ps/Gordon01.pdf]

rocha@lanl.gov

2. Latent Corpora of Biological Data

Active Databases

- Latent Corpura discovers implicit, higher-order associations among stored objects
  - Latent Semantic Analysis
  - Analysis of Graph Structure
    - Links, Distance Functions and Metrics
  - Clustering
  - Works at several levels
    - Within objects, groups of objects, and the entire corpus
- In Information Retrieval latent associations are extracted from the relation between documents and keyterms
  - For biological DKS documents are subtituted by objects with different *types* of tags, leading to many relations to be analysed.
- Relation To Machine Learning Component



Luis Rocha
2001

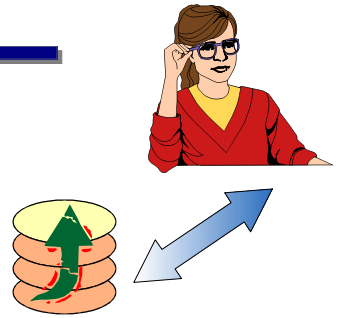
http://www.c3.lanl.gov/~rocha[/ps/Gordon01.pdf]

Los Alamos National Laboratory

rocha@lanl.gov

3. Collaborative and Recommendation Systems

Adaptive and Collective Behavior

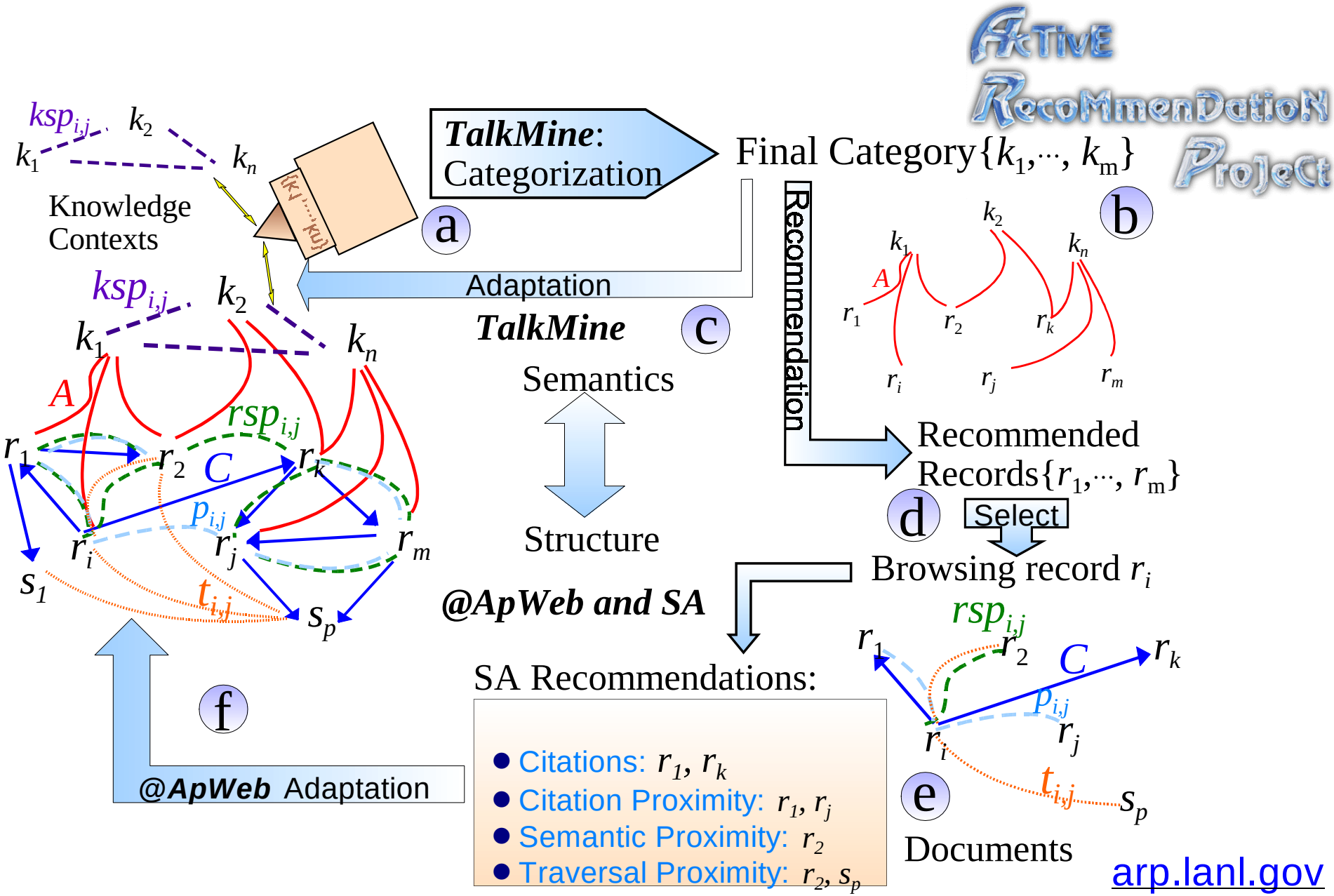


- Tasks of complex biological problems are tackled by large teams and communities.
  - The behavior of these communities can itself be harvested to discover associations between data-sets, hypothesis, etc.
- Recommendation systems use the collective behavior of users (plus latent relations) to discover, categorize, and recommend resources and fellow researchers.

Luis Rocha
2001

http://www.c3.lanl.gov/~rocha[/ps/Gordon01.pdf]

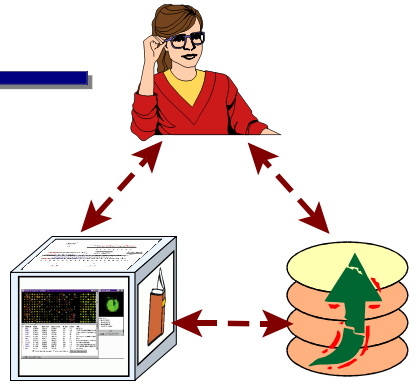
Los Alamos National Laboratory


ACTIVE RECOMMENDATION PROJECT
TalkMine: Categorization
Final Category{k1,···, km}
Knowledge Contexts
a
b
Adaptation
TalkMine
c
Semantics
Structure
Recommendation
Recommended Records{r1,···, rm}
d
Select
Browsing record ri
@ApWeb and SA
SA Recommendations:
Citations: r1, rk
Citation Proximity: r1, rj
Semantic Proximity: r2
Traversal Proximity: r2, sp
f
@ApWeb Adaptation
e
Documents
arp.lanl.gov

rocha@lanl.gov

4. Discovery of Meta-Associations

Analysis of New Associations



- All 3 previous subcomponents aim at building the cabability of automatic generation of associations:
  - 1. Produces intelligent data containers that keep both author-supplied and automatic associations
  - 2. Produces databases that discover latent associations, distance functions, and reduce dimensionality
  - 3. From collective behavior, associations are produced at all levels.
- However, this automatic generation of associations should be itself harvested
  - For generating hypothesis about biological processes to help generate new experiments
  - Discovery of communities of interest
  - DKS as research tools in addition to information retrieval and recommendation systems.
- Relation to Machine Learning Component

Luis Rocha
2001

http://www.c3.lanl.gov/~rocha[/ps/Gordon01.pdf]