Biological Data Management, part 2

H. V. Jagadish
University of Michigan

Outline

- Introduction to Biology and Bioinformatics
- Case Study of a Biological Data Management System
- Technical Challenges
  - Provenance
  - Ontology
  - Usability

Biological ontologies

- Tend NOT to be formal ontologies
- "Practical" ontologies?
- Controlled/structured vocabularies

Biological ontologies

- GO
  - Genome annotation
- MGED
  - Functional genomics experiments
- UMLS
  - "Uber" ontology of ontologies
  - Complete description of medical knowledge

OBO ontologies

- Open and free for use
- Semantic-free unique identifier
  - GO:0006260
- Text definition w/ citation
- Common syntax
  - OBO format
- Orthologonal
  - Over 40 ontologies at obo.sourceforge.net

Creating the Gene Ontology Resource: Design and Implementation

The Gene Ontology Consortium[2]

GO

- Scope: Ontology for gene annotation
  - Species neutral
    - Currently biased towards eukaryotic model organisms
- Source
  - Flybase, Yeast, Mouse
  - Textbooks. Eg. Oxford dictionary of molecular biology
- 18,000+ terms
  - Most terms can be used directly for gene annotation

[Term]
id: GO:0006260
name: DNA replication
namespace: biological_process
def: "The process whereby new strands of DNA are synthesized. The template for replication can either be DNA or RNA." [ISBN:0198506732]
comment: See also the biological process terms 'DNA-dependent DNA replication ; GO:0006261' and 'RNA-dependent DNA replication ; GO:0006278'.
subset: gosubset_prok
synonym: "DNA biosynthesis"
synonym: "DNA replication accessory factor"
synonym: "DNA replication factor"
synonym: "DNA synthesis"
is_a: GO:0006259 ! DNA metabolism

GO divisions

- Molecular Function
  - Enzyme, transporter, …
- Biological process
  - Signal transduction, fatty acid metabolism, …
- Cellular component
  - Location in the cell, nuclear membrane

Annotating with GO

- Assignments are independent
  - Genes have multiple functions
  - Function does not infer process
- Annotations must have supporting evidence
- Evidence code + external cross refrence
  - IC: Inferred by Curator
  - IDA: Inferred from Direct Assay
  - IEA: Inferred from Electronic Annotation
  - IEP: Inferred from Expression Pattern
  - IGI: Inferred from Genetic Interaction
  - IMP: Inferred from Mutant Phenotype
  - IPI: Inferred from Physical Interaction
  - ISS: Inferred from Sequence or Structural Similarity
  - NAS: Non-traceable Author Statement
  - ND: No biological Data available
  - RCA: inferred from Reviewed Computational Analysis
  - TAS: Traceable Author Statement
  - NR: Not Recorded
- Provides hint of annotation quality!

The MGED Ontology: a resource for semantics-based description of microarray experiments

Patricia L. Whetzel[1], Helen Parkinson[2], Helen C. Causton[3], Liju Fan[4], Jennifer Fostel[5], Gilberto Fragoso[6], Laurence Game[3], Mervi Heiskanen[6], Norman Morrison[7], Philippe Rocca-Serra[2], Susanna-Assunta Sansone[2], Chris Taylor[2], Joseph White[8] and Christian J. Stoeckert, Jr[1,*]

MGED Ontology

- MGED Ontology (MO) and MGED Core Ontology (MCO)
- All aspects of a microarray experiment
  - Experimental design, sample preparation, assay and analysis protocols
- 229 classes, 110 properties, 658 instances
- http://mged.sourceforge.net/ontologies/MGEDontology.php

Design

- Classes/concepts
- Attributes/properties
- Actual values/instances

- Supports the MAGE object model

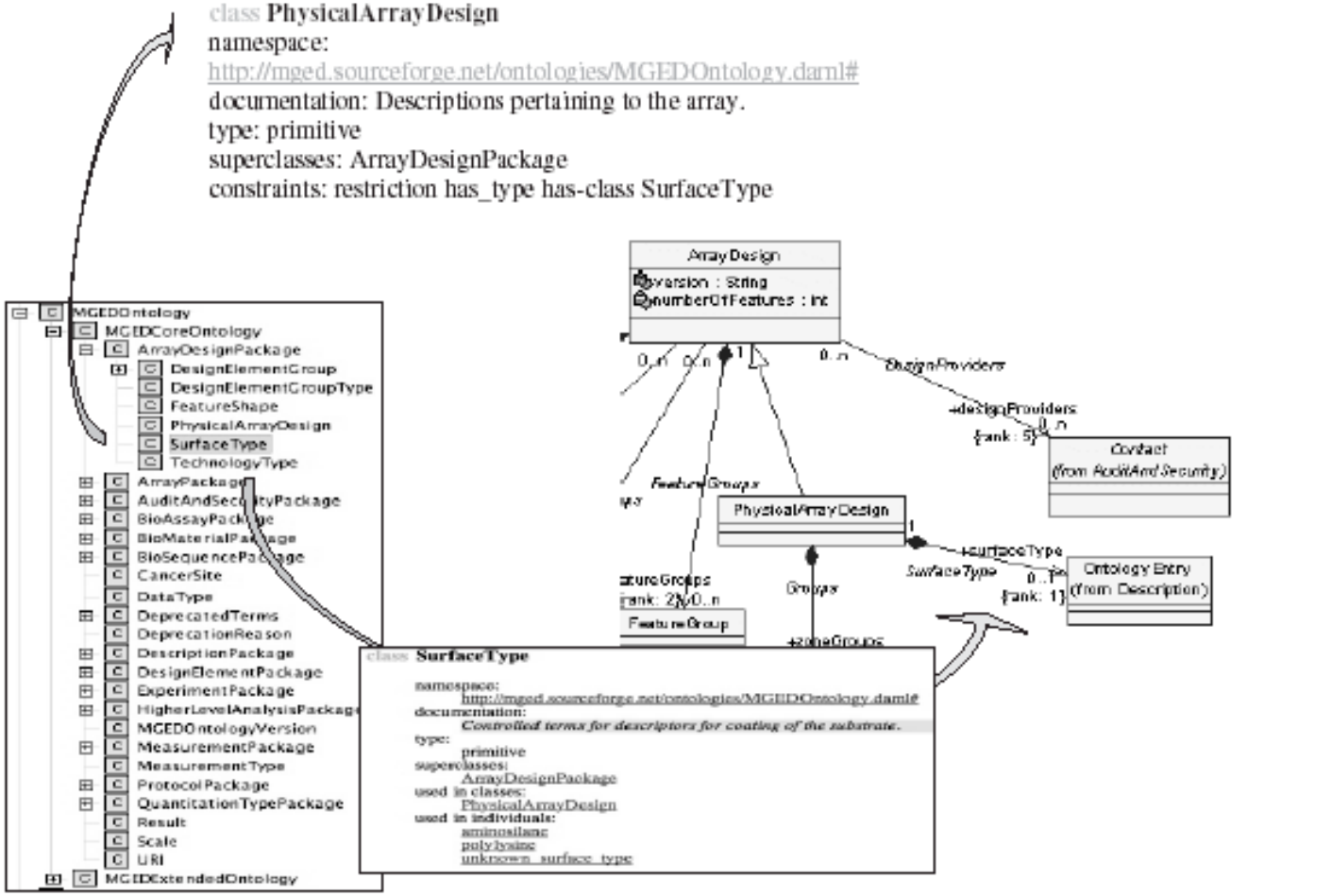


Fig. 2. Class hierarchy of the MO and relationship to the MAGE-OM. In this example, the MAGE-OM specifies a 'surfaceType' association to OntologyEntry from PhysicalArrayDesign. Terms (polylysine, aminosilane, unknown_surface_type) for surface type can be found in the MO in the class 'SurfaceType' which is located in the ArrayDesignPackage class. The relationship of SurfaceType to PhysicalArrayDesign is captured in MO: (PhysicalDesignType has_type SurfaceType).

Journal of the American Medical Informatics Association Volume 5 Number 1 Jan / Feb 1998

Focus on **The UMLS**

JAMIA

Technical Milestone ■

The Unified Medical Language System:

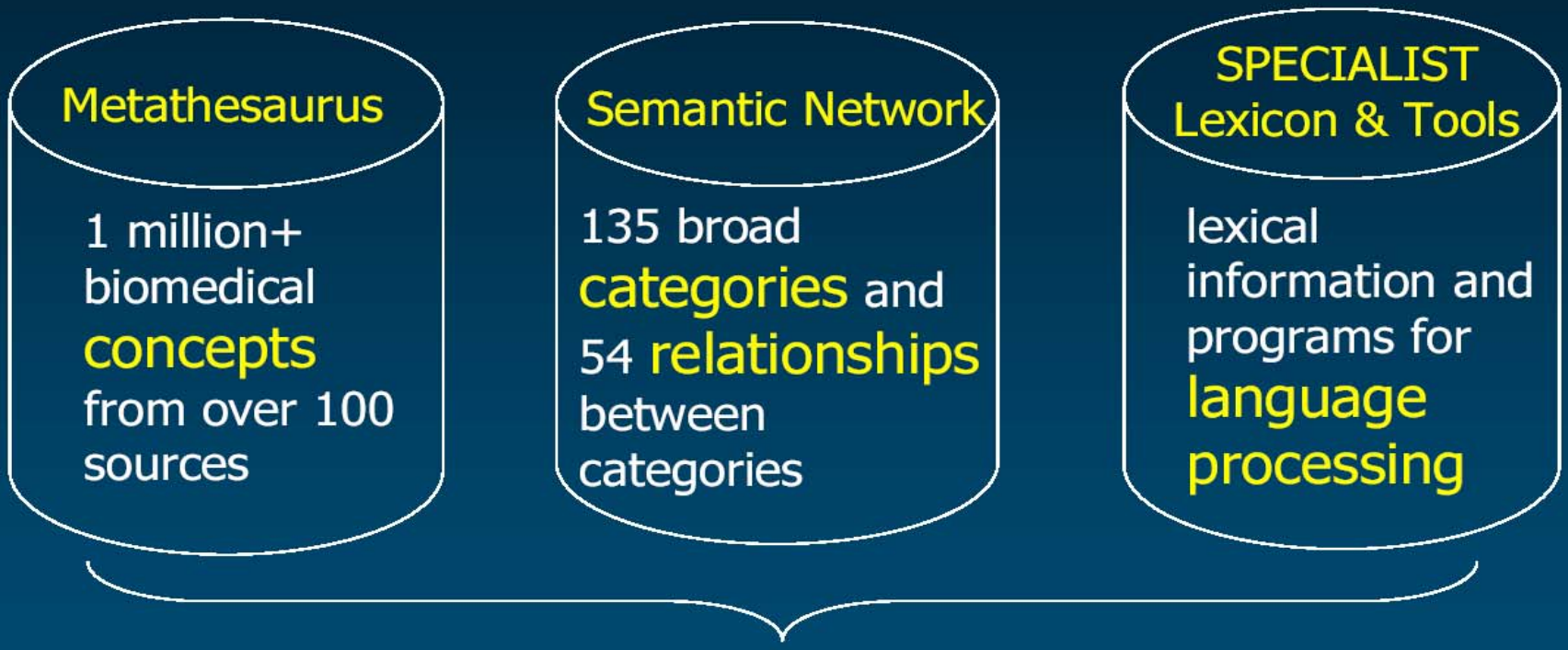
An Informatics Research Collaboration

Betsy L. Humphreys, MLS, Donald A. B. Lindberg, MD,
Harold M. Schoolman, MD, G. Octo Barnett, MD

Motivation

- "the principal barrier to effective integrated access to biomedical information is the tremendous array of classification …the solution to this fundamental medical information problem is the development of conceptual links among disparate classification schemes...."
  - UMLS RFP 1986

The UMLS consists of



3 Knowledge Sources
used separately or together

Slides reproduced from http://www.nlm.nih.gov/research/umls/pdf/**UMLS_Basics**.pdf

Metathasaurus

- **Enormous**
- combined scope of its 100+ source vocabularies
- Preservation of Content and Meaning from Source Vocabularies
- Customizable, trimmed via software

MESH

- Medical subject headings
  - Anatomy
  - Mental disorders
- 22,997 descriptors
  - Thousands more cross-references/synonyms
- Manually collected from literature
- Used to index MEDLINE/PubMED entries

ICD

- International Statistical Classification of Diseases and Related Health Problems
- Coding system for diseases
- Developed by WHO starting in 1948
- 10th major edition.
  - 3 yearly updates
- (A05.) Other bacterial foodborne intoxications
  - (A05.0) Foodborne staphylococcal intoxication

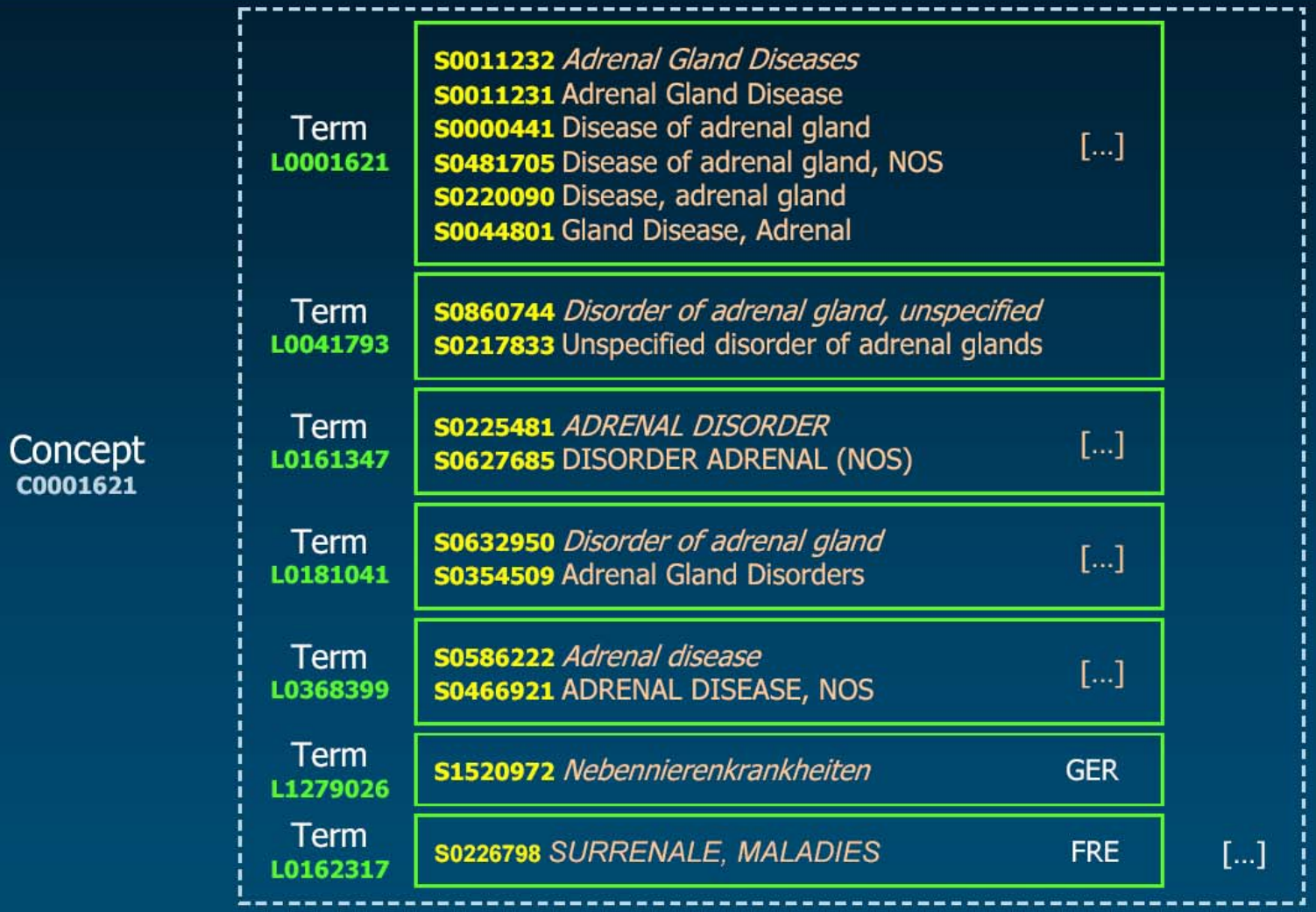
Metathesaurus: clusters terms by meaning

- Synonymous terms clustered into a concept
- Preferred term is chosen
- Unique identifier (CUI) is assigned

Addison's disease	Metathesaurus	PN	
Addison's disease	SNOMED CT	PT	363732003
Addison's Disease	MedlinePlus	PT	T1233
Addison Disease	MeSH	PT	D000224
Bronzed disease	SNOMED Intl 1998	SY	DB-70620
Deficiency; corticorenal, primary	ICPC2-ICD10 Thesaurus	PT	MTHU021575
Primary Adrenal Insufficiency	MeSH	EN	D000224
Primary hypoadreanlism syndrome, Addison	MedDRA	LT	10036696

C0001403 Addison's disease

Cluster of synonymous terms



Outline

- Introduction to Biology and Bioinformatics
- Case Study of a Biological Data Management System
- Technical Challenges
  - Provenance
  - Ontology
  - Usability
  - http://www.eecs.umich.edu/db/usable
  - H. V. Jagadish et al, "Making Database Systems Usable," SIGMOD 2007.

Obvious Challenges

- Unknown Query Language
- Unknown Schema
- Complex Schema
- Unknown Data Values

Challenge: Unknown Query Language

```
for   $a in doc()//author,
      $s in doc()
let   $
wh
```

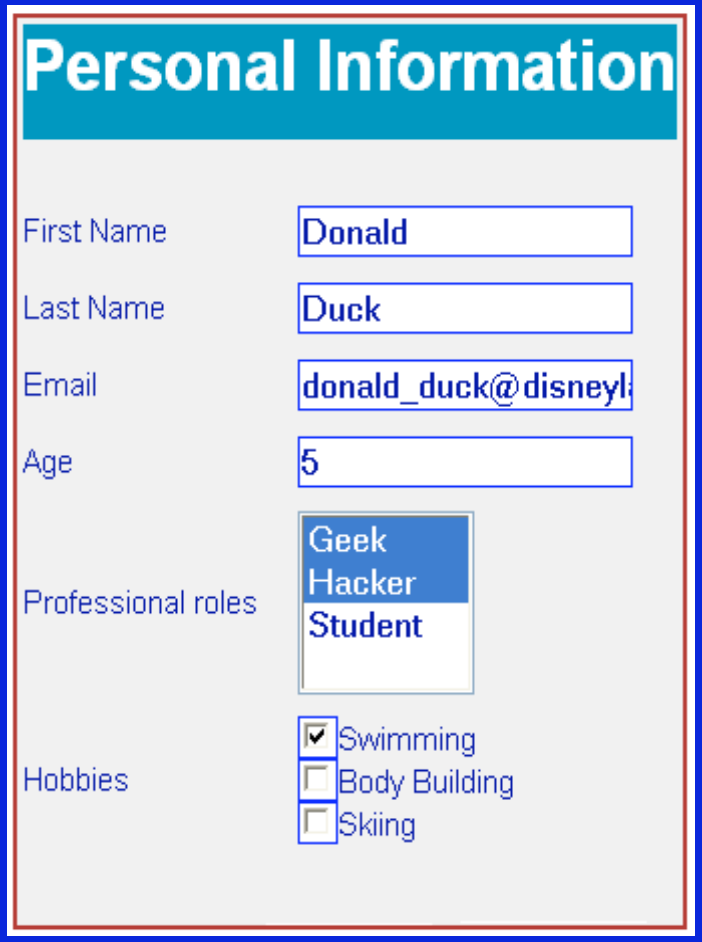
$a ??
What is *let*?
Do I need a semi-colon?
How do I start writing a query?

Challenge: Unknown Query Language

- Solutions:
  - Forms
  - Natural Language Query

Forms: Magesh Jayapandian

- Simple, but limited.
- How to create a good set of query forms?
- Can we let a user modify a form that "almost" does the desired thing?



Natural Language Query: Yunyao Li

- A generic interface supporting English queries to a database.
- Follow Up Queries: conversational iterative specification of queries.
- Add Domain Knowledge learning component to improve the generic interface.

Challenges in Natural Language Querying

- *Challenge 1:*

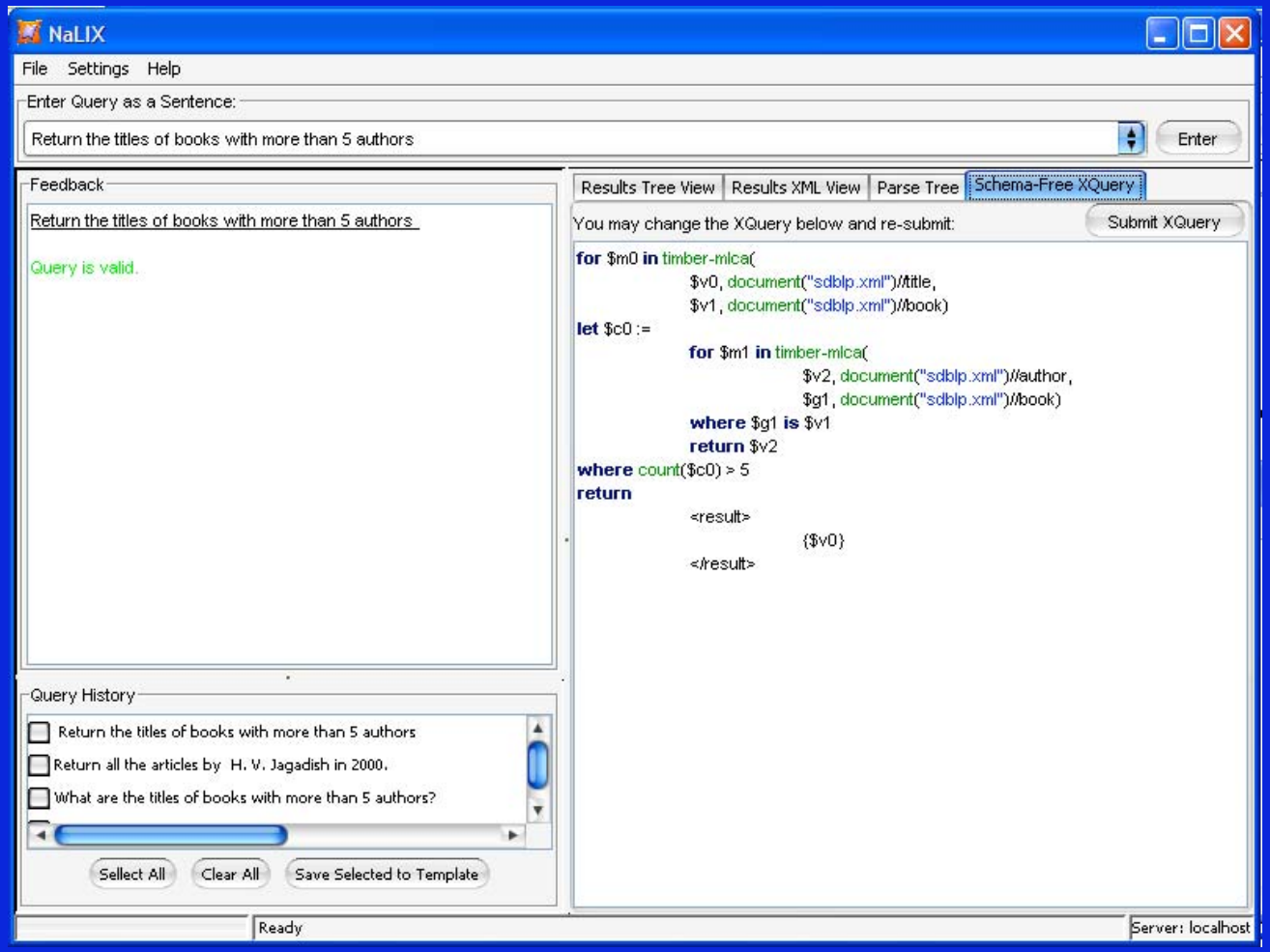
  Understand user intent given an arbitrary natural language query.

- *Challenge 2:*

  Map user intent to database schema.

  - Is "Gone with the wind" a book or a movie (or a person)?
  - Are books grouped by year or by author in the bibliography?

Example – Nesting



Q: Return the titles of books with more than 5 authors.

Challenge: Unknown Schema

Aaron Elkiss, Yunyao Li, Cong Yu


```
for   $a
      $s
let   $b
where
  $s/co                    on" and
    $b/a
return {
```

Schema-Free XQuery

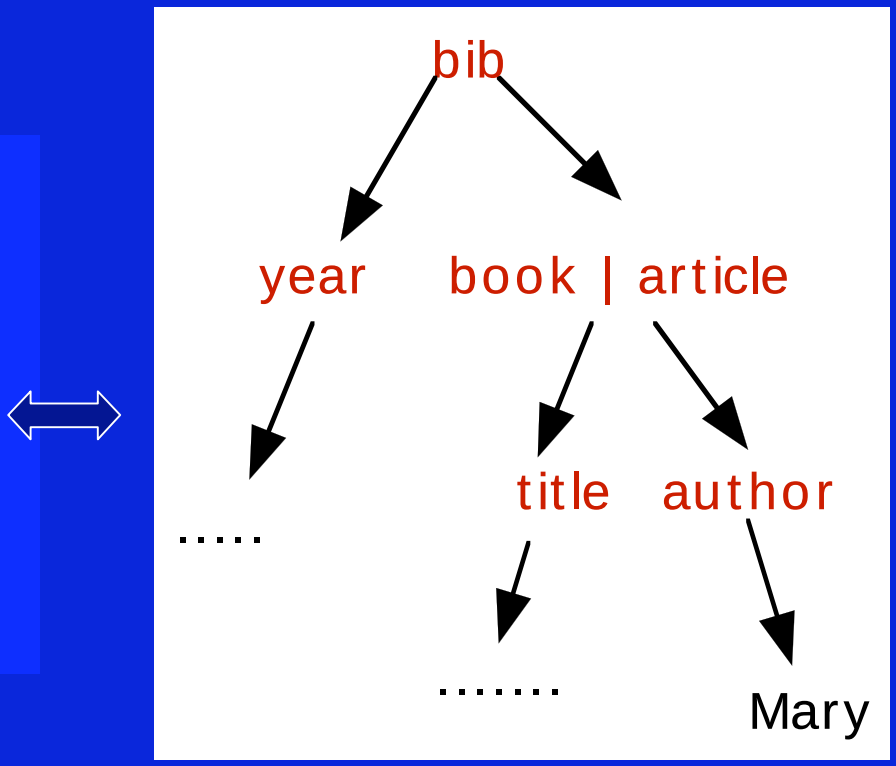
Enable users to query XML data by exploiting whatever partial knowledge of the schema they have: support wide range of queries - from regular XQuery to keyword search.

Extended from Boolean notion of correctness to a notion of "ranked relatedness", permitting seamless transition to IR-style querying.

Traditional Query Focus

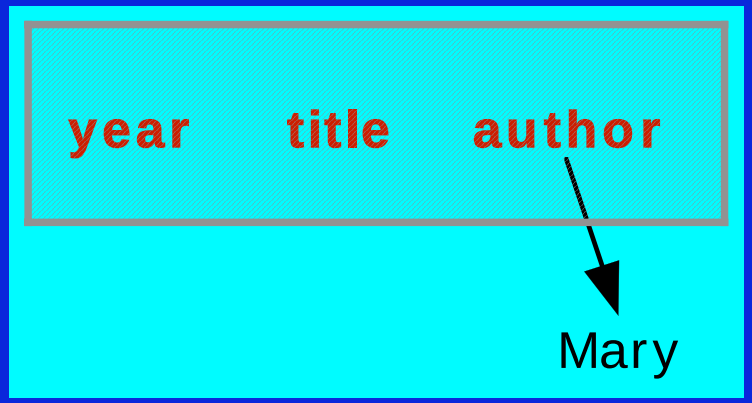
- Knowing the document structure, the user can specify in XQuery **HOW** the nodes are related in terms of structural relationship:

```
for $b in doc("bib.xml")/bib
for $c in $b/book or $b/article
where $c/author = "Mary"
return {
  <result>
    $c/title
    $b/year
  </result>
}
```

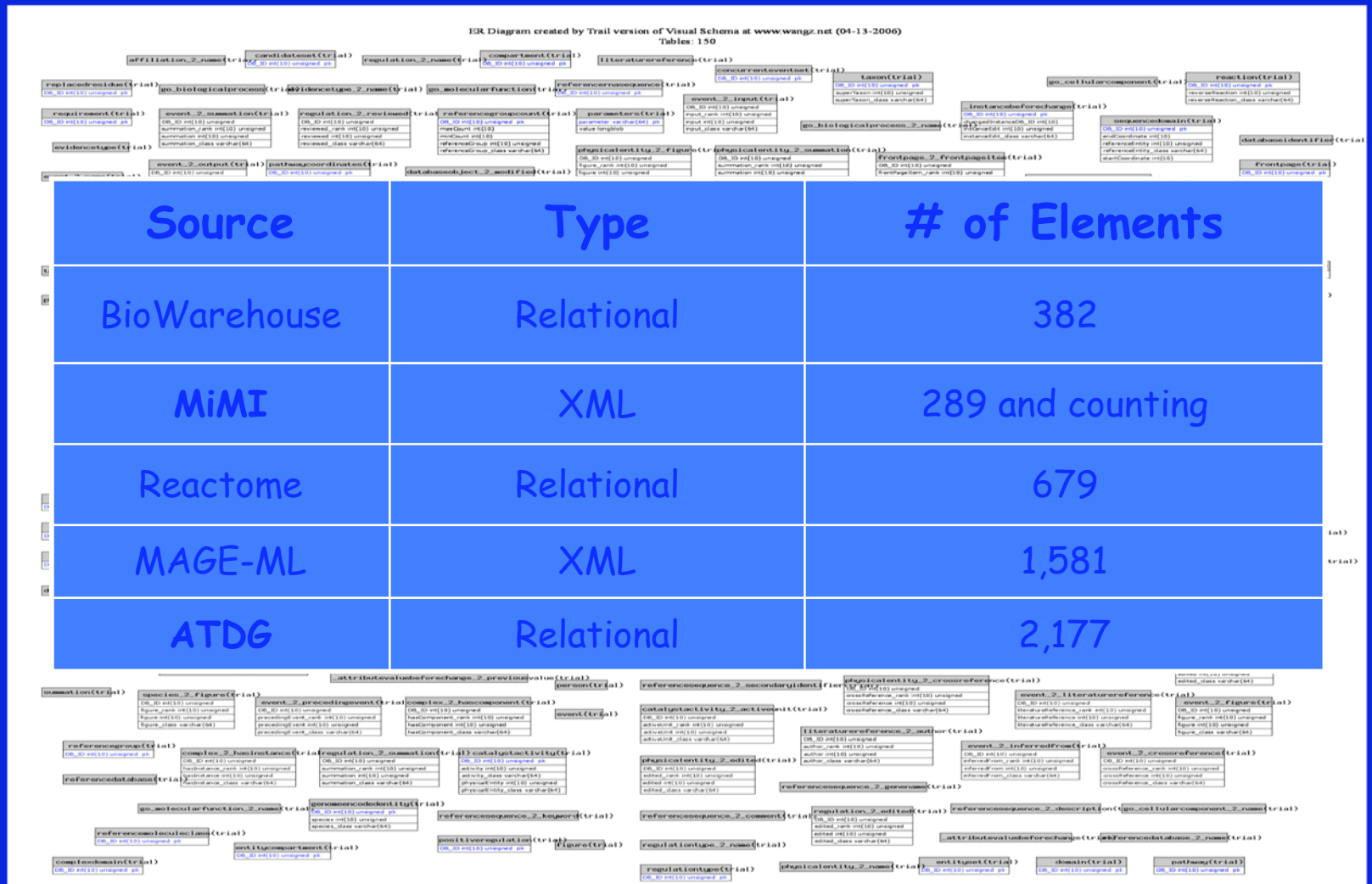


Schema-Free Query Focus

- Without knowing the document structure, the user can still specify **WHICH** nodes should be meaningfully related:



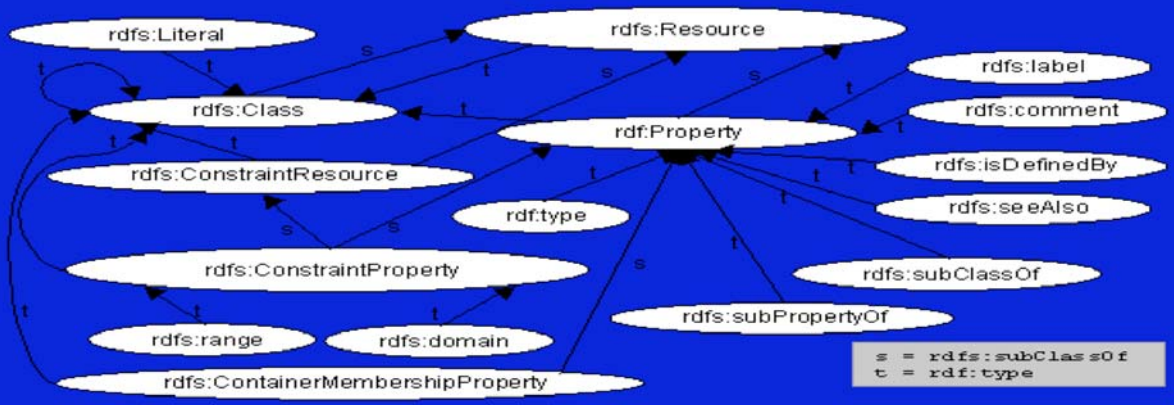
Challenge: Complex Schema



Source	Type	# of Elements
BioWarehouse	Relational	382
MiMI	XML	289 and counting
Reactome	Relational	679
MAGE-ML	XML	1,581
ATDG	Relational	2,177

Schema Summarization: Cong Yu

- Schema are often too large and too complex.



- Can we present the user with an informative summary?
- Can the user effectively query the database using this summary alone?

Schema Summarization

- Basic Idea:
  - Represent the original complex schema with a smaller and conceptually simpler schema – a *summary* of the original schema.
  - Each element in the summary naturally corresponds to a *subschema* of the original schema.
- Helps users *explore* the schema:
  - Illustrates the main topics of the database.
  - Filters away irrelevant parts of the schema.

Schema Summary

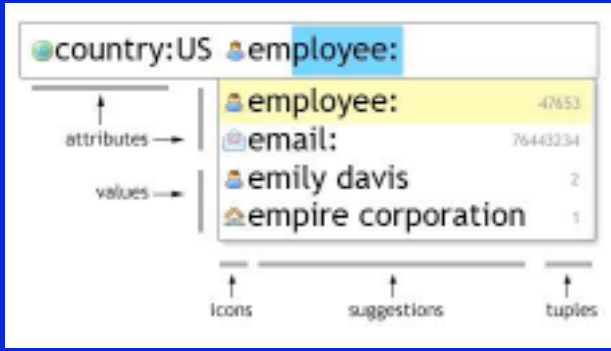
- Summary is a schema:
  - Contains *abstract elements* and *abstract links;*
  - Smaller in size.
- Abstract element:
  - Represents a subschema, i.e., a group of original elements.
- Abstract link:
  - Connects abstract elements.

Challenge: Unknown Data Values


```
for   $a
      $s
let   $b
where
  $s/co                    on" and
   $b/a
return {
```

Autocompletion: Arnab Nandi

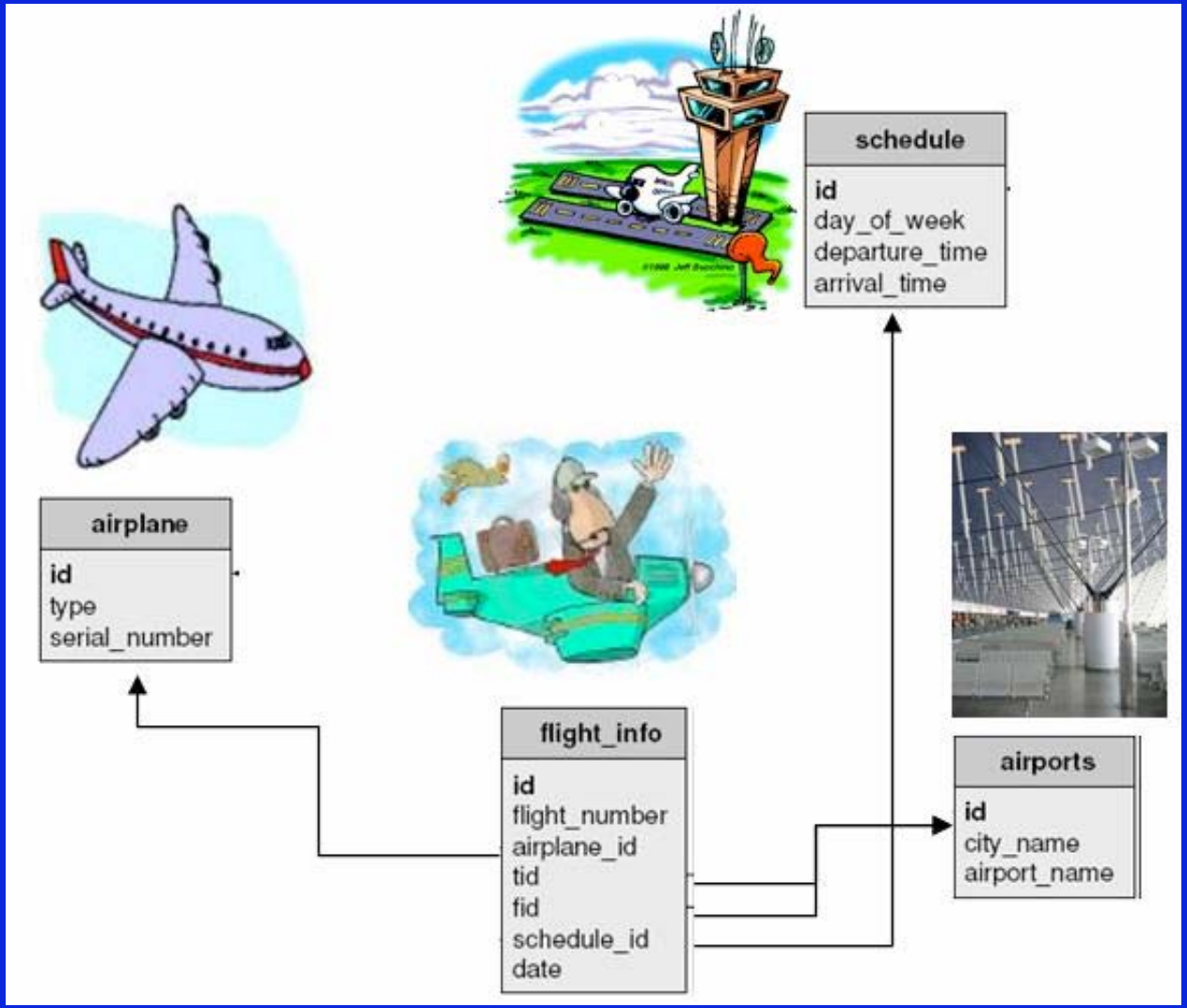
- Help the user along with "instant" feedback as they type.
- Provide insights into schema, data and familiar syntax *during* query formulation.
- Guide them to perform better queries, correctly.

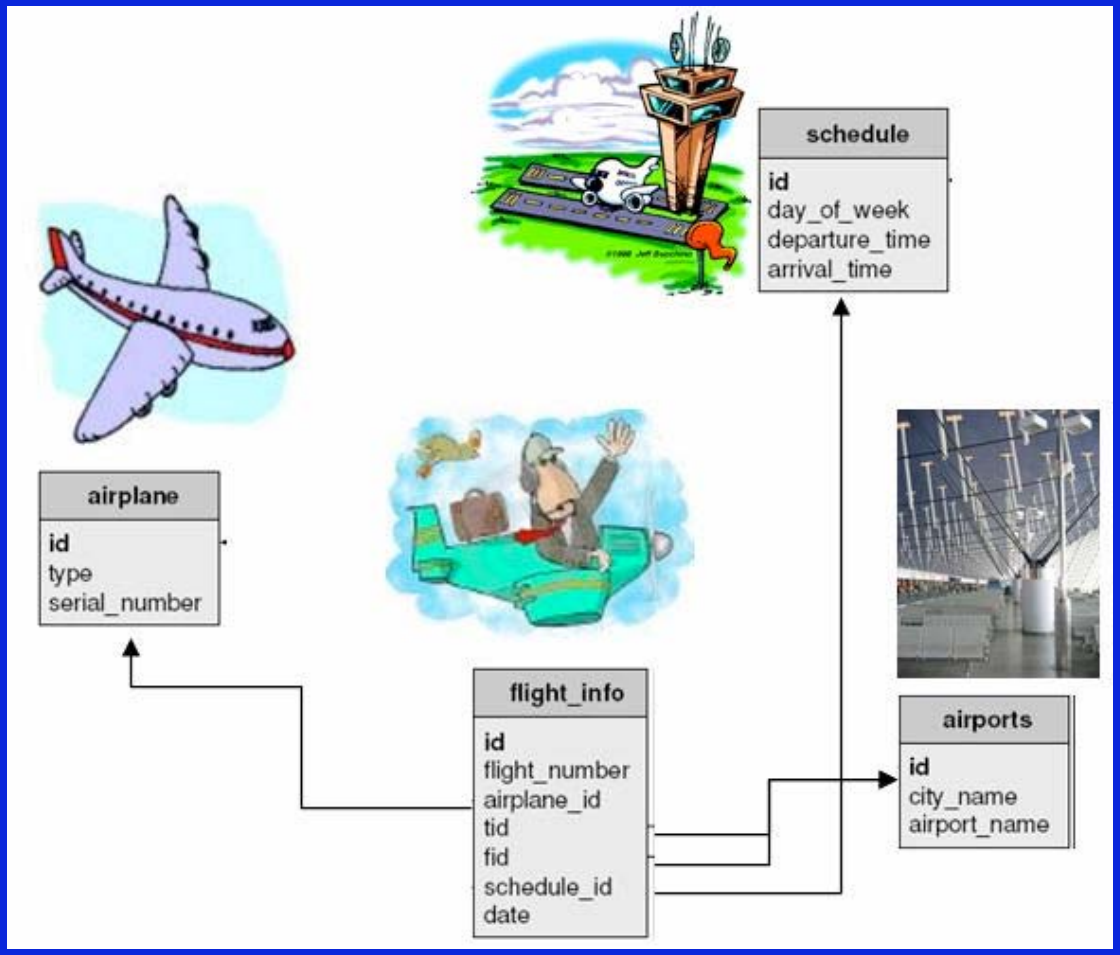


Deeper Challenges

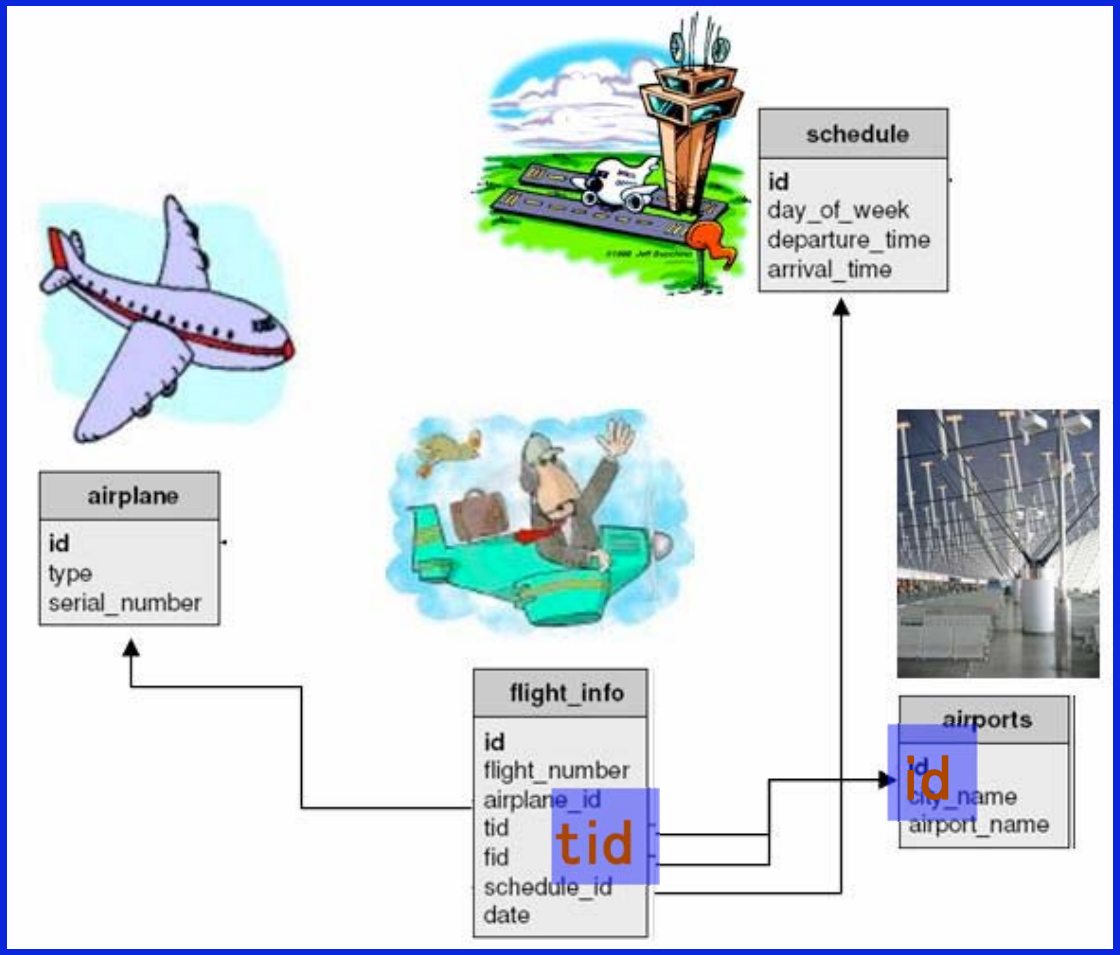
- Too many joins
- Too many options
- No direct manipulation

Painful Relations





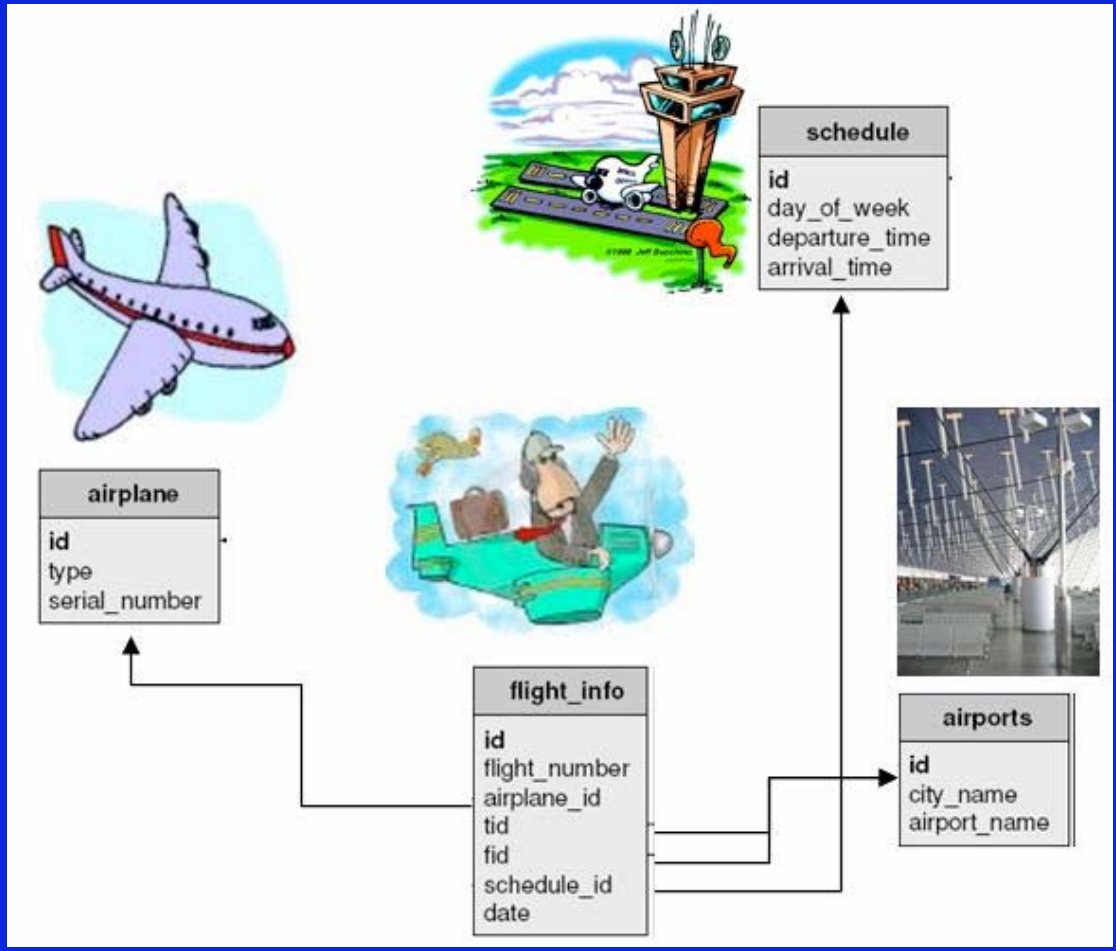
Single user concept (Flight) has been normalized into four tables.



Names of tables and attributes are not self-explanatory, particularly where references are involved (fid, tid).

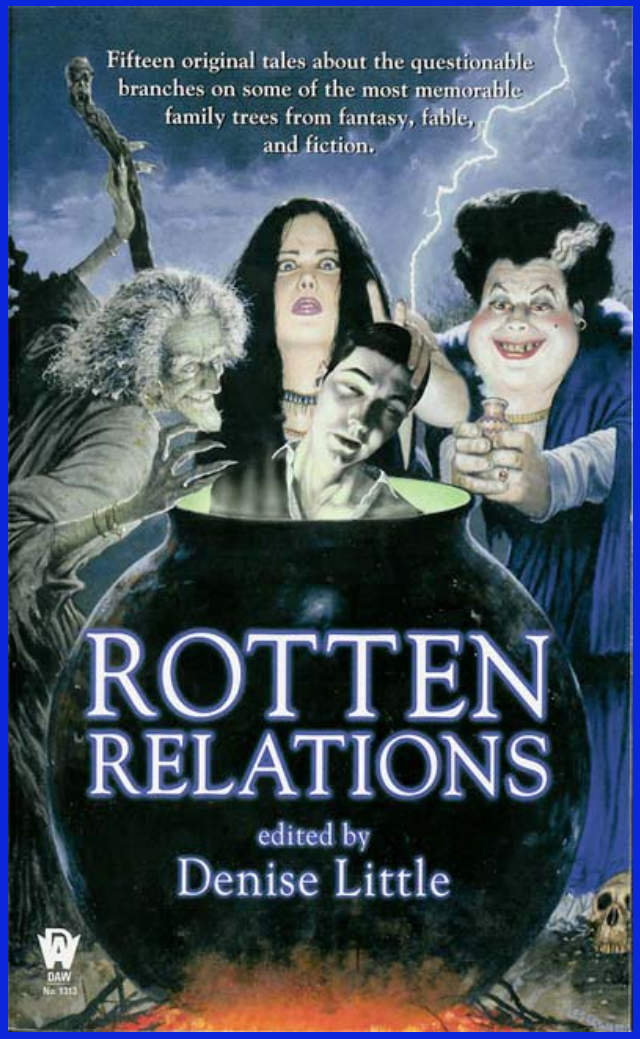
Find departure times for flights from Beijing to Detroit.

```
SELECT s.departure_time
FROM schedule s,
   flight_info f, airports d,
   airports a
WHERE s.id = f.schedule_id
 AND f.fid = d.id
 AND d.city_name = “Beijing”
 AND f.tid = a.id
 AND a.city_name = “Detroit”
```



Even simple queries are not easy to express.

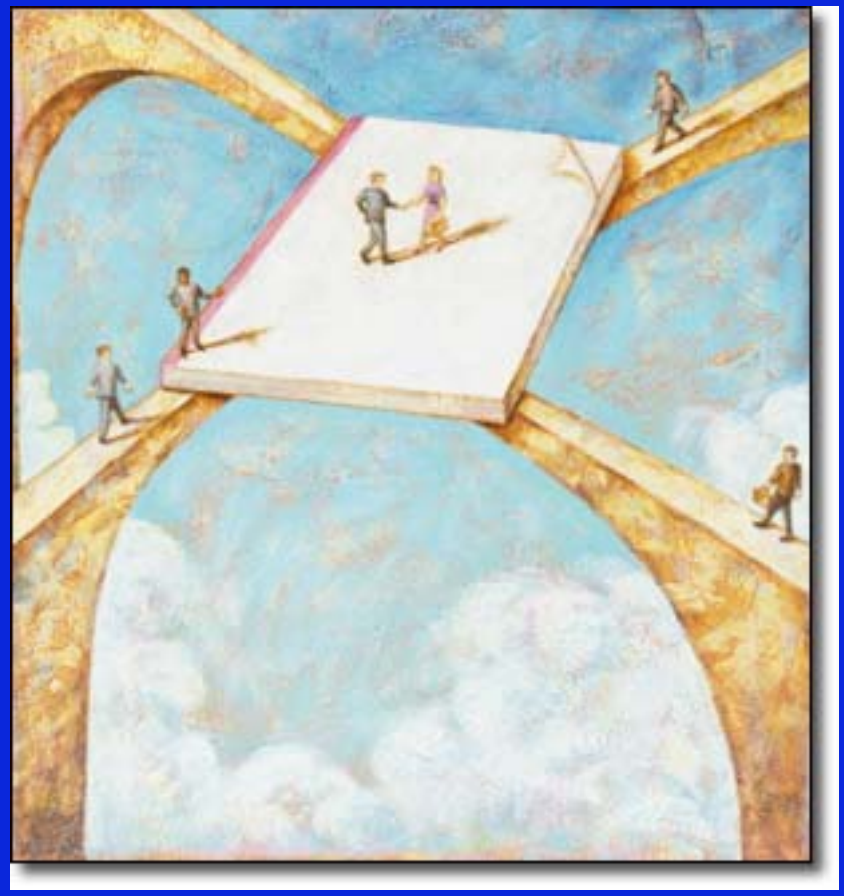
Not Just Relations!



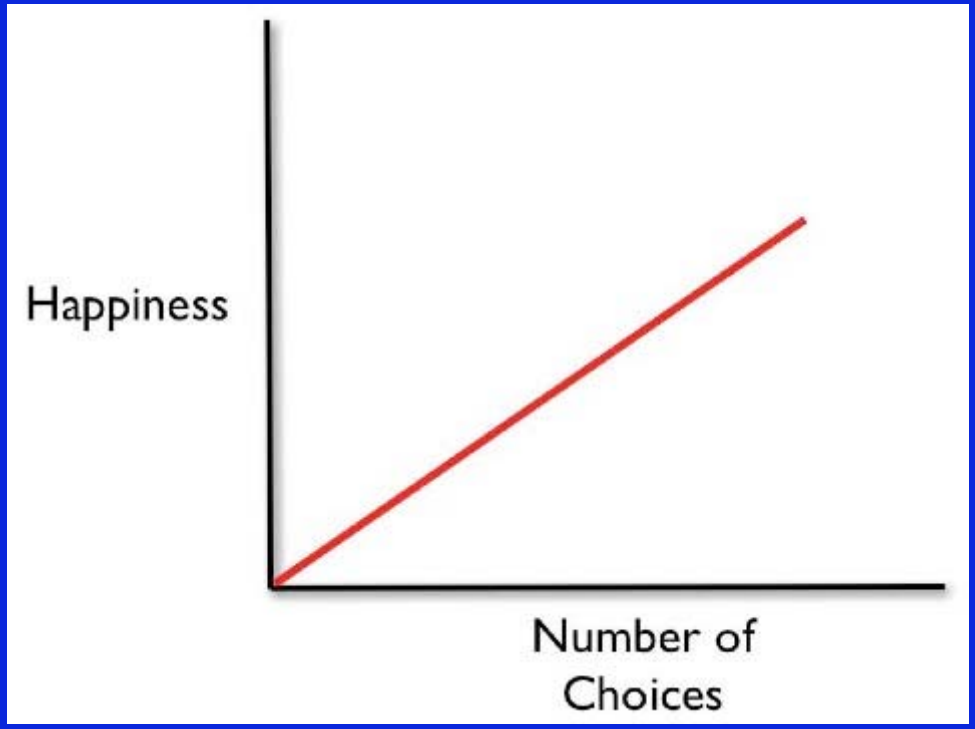
- Relational value joins may be the worst offender.
- But XML joins are bad too:
  - ID/IDREF
  - Structural

1. No Joins

The typical user will only be able to express selection/projection: no joins.

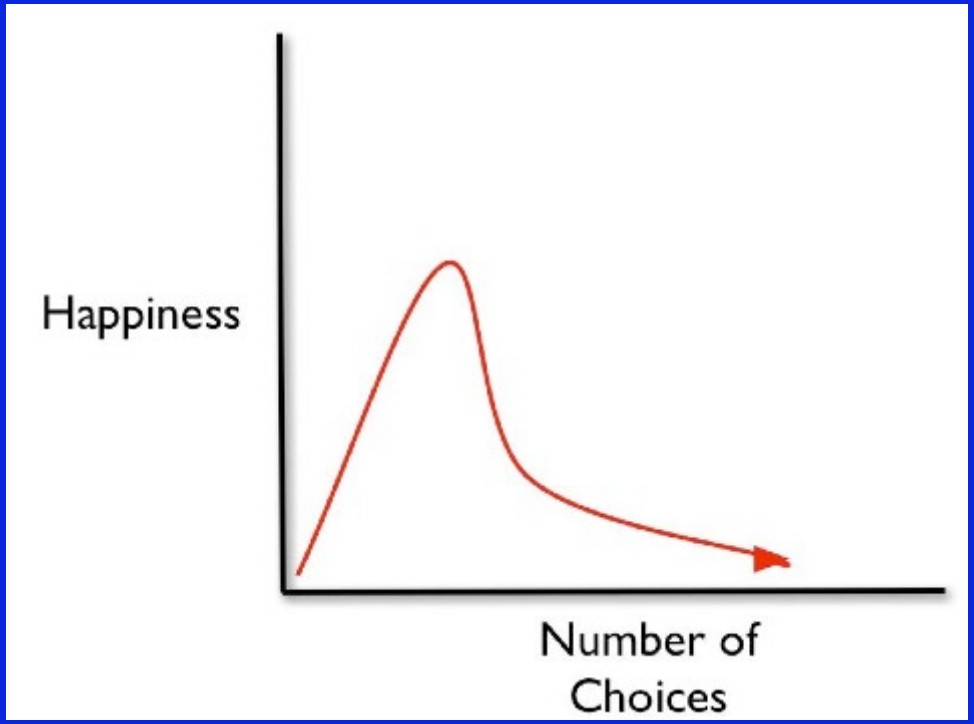


Painful Options



What a software designer thinks is true

The Fallacy of Greater Choice



Barry Schwartz, The tyranny of choice. Scientific American, April 2004, pp. 71-75

2. Limited Options

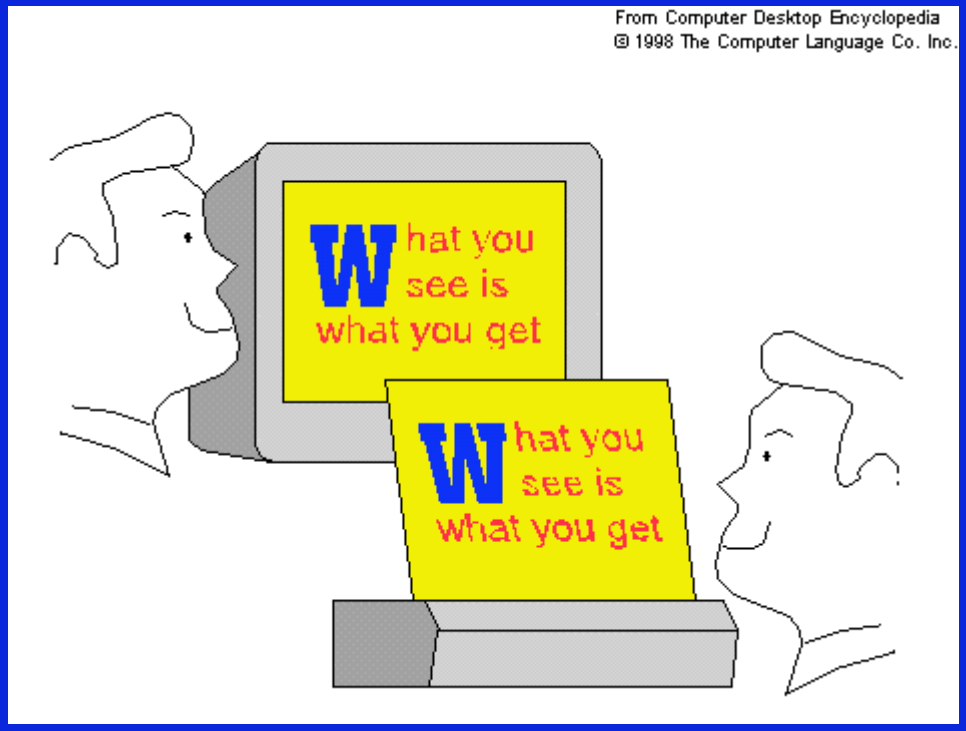
An ideal system will provide just enough options for the user to get their work done, but no more.

Or provide a gradual migration path with more options for the more advanced user.

Invisible Pain

Which Word Processor Do You Use?

If, like me, you said LaTeX, then you are not a typical user.

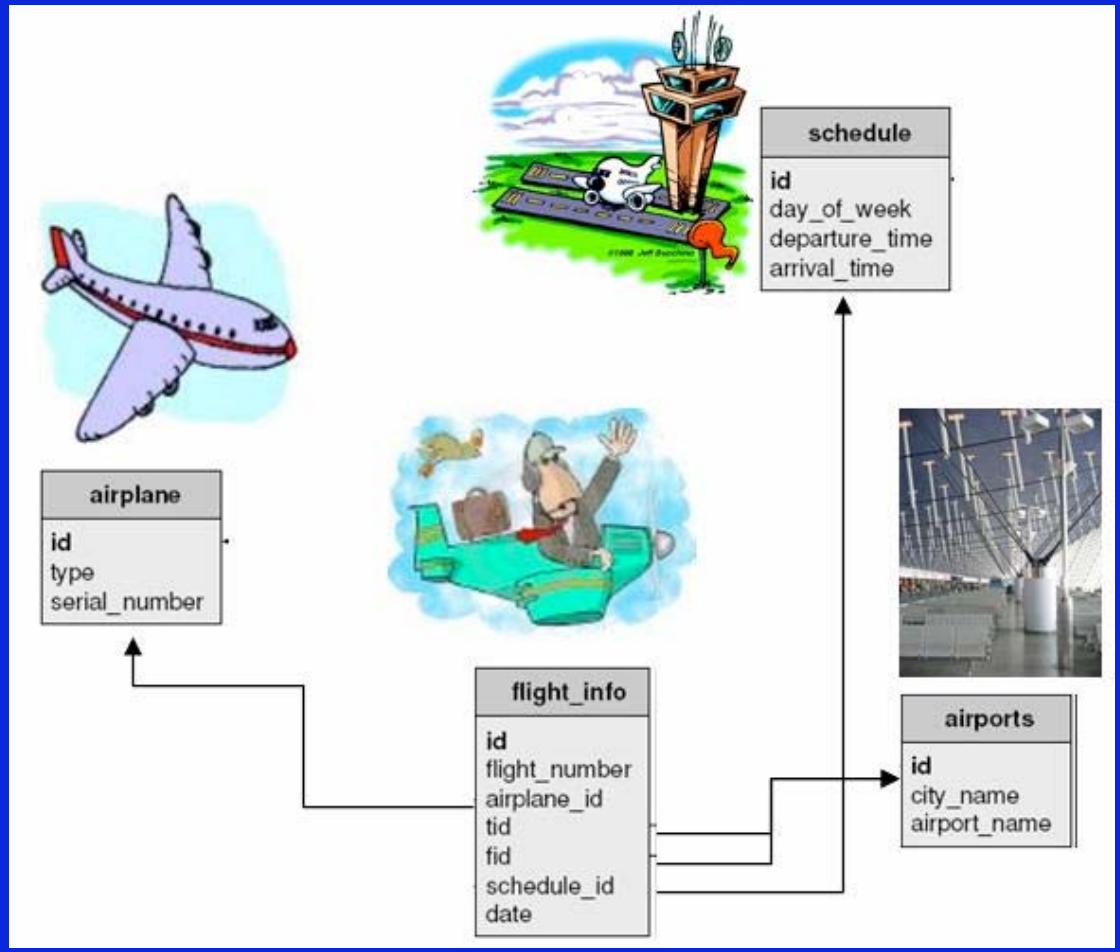


Very hard to specify changes in the abstract, programmatically.

Much easier to work with the concrete: click and drag and drop.

Find departure times for flights from Beijing to Detroit.

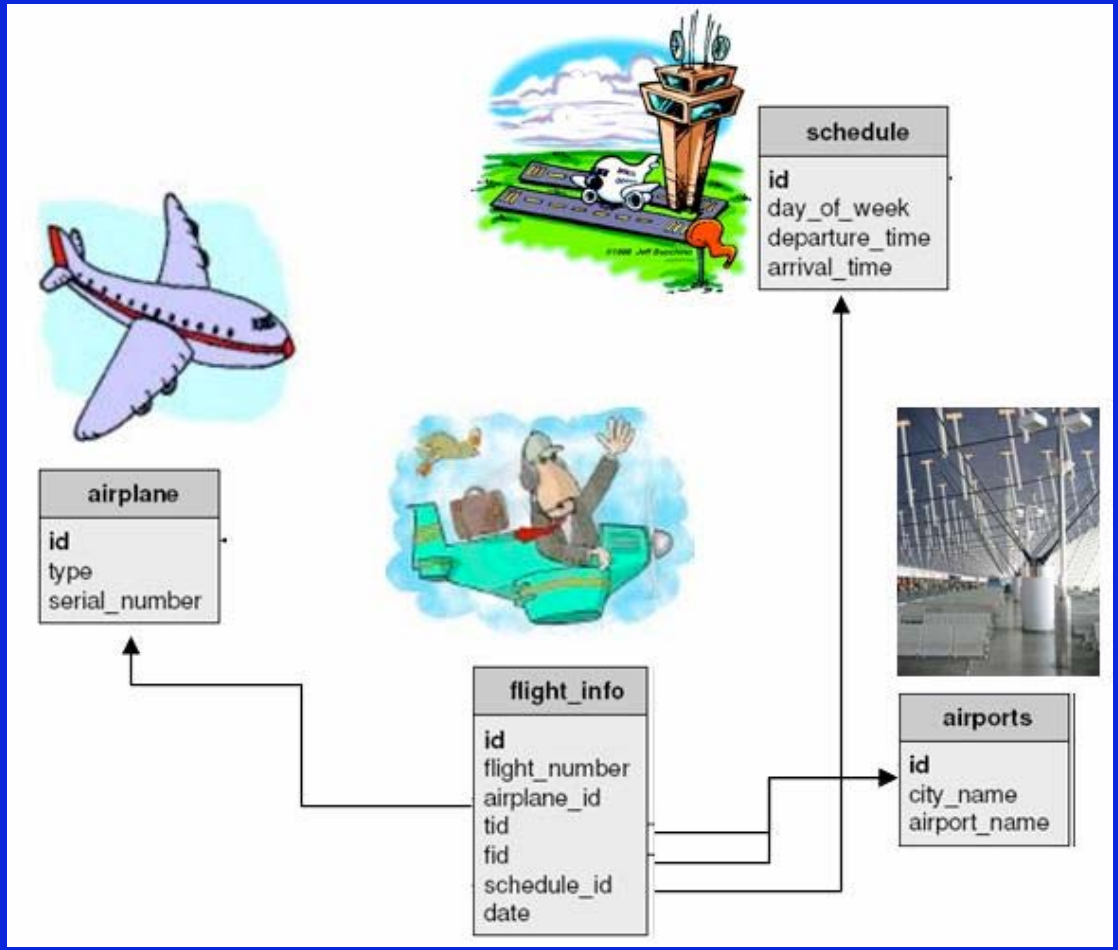
```
SELECT s.departure_time
FROM schedule s,
   flight_info f, airports d,
   airports a
WHERE s.id = f.schedule_id
 AND f.fid = d.id
 AND d.city_name = “Beijing”
 AND f.tid = a.id
 AND a.city_name = “Detroit”
```



Even small changes can be difficult to make.

Find departure times for 747 flights from Beijing to Detroit.

```
SELECT s.departure_time
FROM schedule s,
   flight_info f, airports d,
   airports a, airplane p
WHERE s.id = f.schedule_id
 AND f.fid = d.id
 AND d.city_name = “Beijing”
 AND f.tid = a.id
 AND a.city_name = “Detroit”
 AND f.airplane_id = p.id
 AND p.type = “747”
```



3. Direct Manipulation

- Do not expect users to write queries in one window and see results in another.
  - Even most visual query builders require abstraction.
- Allow users to specify the queries iteratively by manipulating the "current" (intermediate) result set shown.

Desiderata

1. No Joins
2. Limited Options
3. Direct Manipulation

Presentation Data Model

- The logical data model provides physical data independence.
  - User does not have to worry about *indices*, *file structure*, *access methods*, …
- The presentation data model provides logical data independence.
  - User does not have to worry about *relations*, *joins*, *keys*, *SQL*, …
  - A conceptually simple view of database.

Presentation Data Model

Presentation

Logical

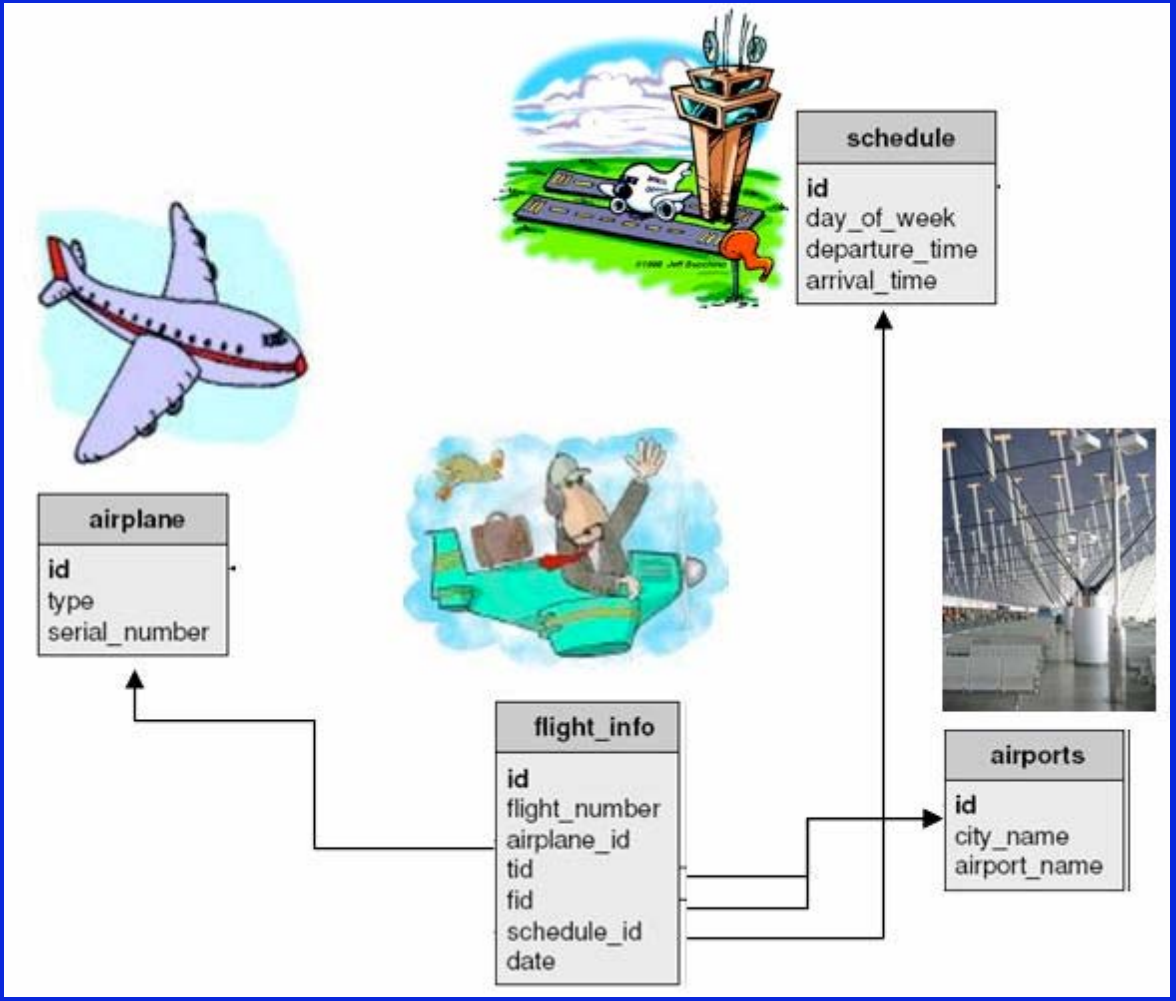
Physical

Data Model + Algebra Layer

Data Model + Algebra Layer

Data Model + Algebra Layer

Flights Database Logical Schema



Flights Database Presentation Schema

Flight

Flight Number	Airplane Type	Date	From City	Departure Time	To City	Arrival Time
23	DC9	6/15	Beijing	1030	Detroit	1515
201	747	6/15	Beijing	2230	Detroit	0550
152	A330	6/15	Beijing	0700	Detroit	1145
9	DC10	6/15	Beijing	0900	Detroit	1630

Relieving Pain from Relations

- User queries the concept of *flight* in the presentation schema.
  - No need to understand the underlying joins
  - No need even to know there are joins
  - E.g., "Give me *flights* from Beijing to Detroit, leaving on June 15th afternoon."
- The system translates the presentation level query into the underlying logical query.

Relieving Pain From Options

- The *Flights* "relation" allows far fewer queries (in a join-free manner) than is possible with arbitrary joins over the logical relations.
- User (at most) specifies:
  - Selection predicates;
  - Attributes retained in projection.
- Further restrictions may be appropriate.

Restricted Presentation Model

- The user only has two options:
  - User specifies time and cities
    - Show flights to/from airports around the cities *geographically* on a *map.*
  - User specifies cities
    - Show flights based on a *timeline.*

- *Real example likely to have a few more.*

Relief from Invisible Pain

Given a simple presentation model, it becomes possible to specify direct manipulation of results as new queries.

Flight Number	Airplane Type	Date	From City	Departure Time	To City	Arrival Time
23	DC9	6/15	Beijing	1030	Detroit	1515
201	747	6/15	Beijing	2230	Detroit	0550
152	A330	6/15	Beijing	0700	Detroit	1145
9	DC10	6/15	Beijing	0900	Detroit	1630

Relief from Invisible Pain

Given a simple presentation model, it becomes possible to specify direct manipulation of results as new queries.

Flight Number	Airplane Type	Date	From City	Departure Time	To City	Arrival Time
23	DC9	6/15	Beijing	1030	**Delhi**	1515
201	747	6/15	Beijing	2230	Detroit	0550
152	A330	6/15	Beijing	0700	Detroit	1145
9	DC10	6/15	Beijing	0900	Detroit	1630

Relief from Invisible Pain

Given a simple presentation model, it becomes possible to specify direct manipulation of results as new queries.

Flight Number	Airplane Type	Date	From City	Departure Time	To City	Arrival Time
275	767	6/15	Beijing	1000	Delhi	1345
277	767	6/15	Beijing	1800	Delhi	2150

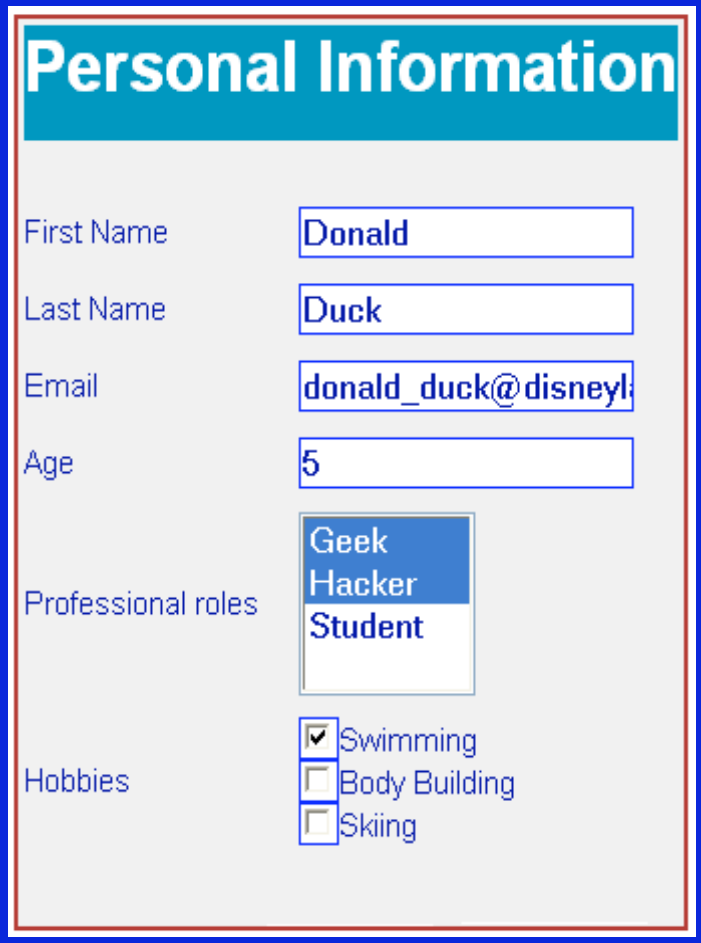
Which systems have this architecture?

- No one in its entirety.
- But

  There are several systems that come close and begin to address some of our requirements.

Forms as Presentation Model

- Provide user with a limited number of useful "views".
- Not perfect:
  - No real model;
  - Little or no explanation;
  - No direct manipulation;
  - No structure creation.
- Yet, wildly popular.



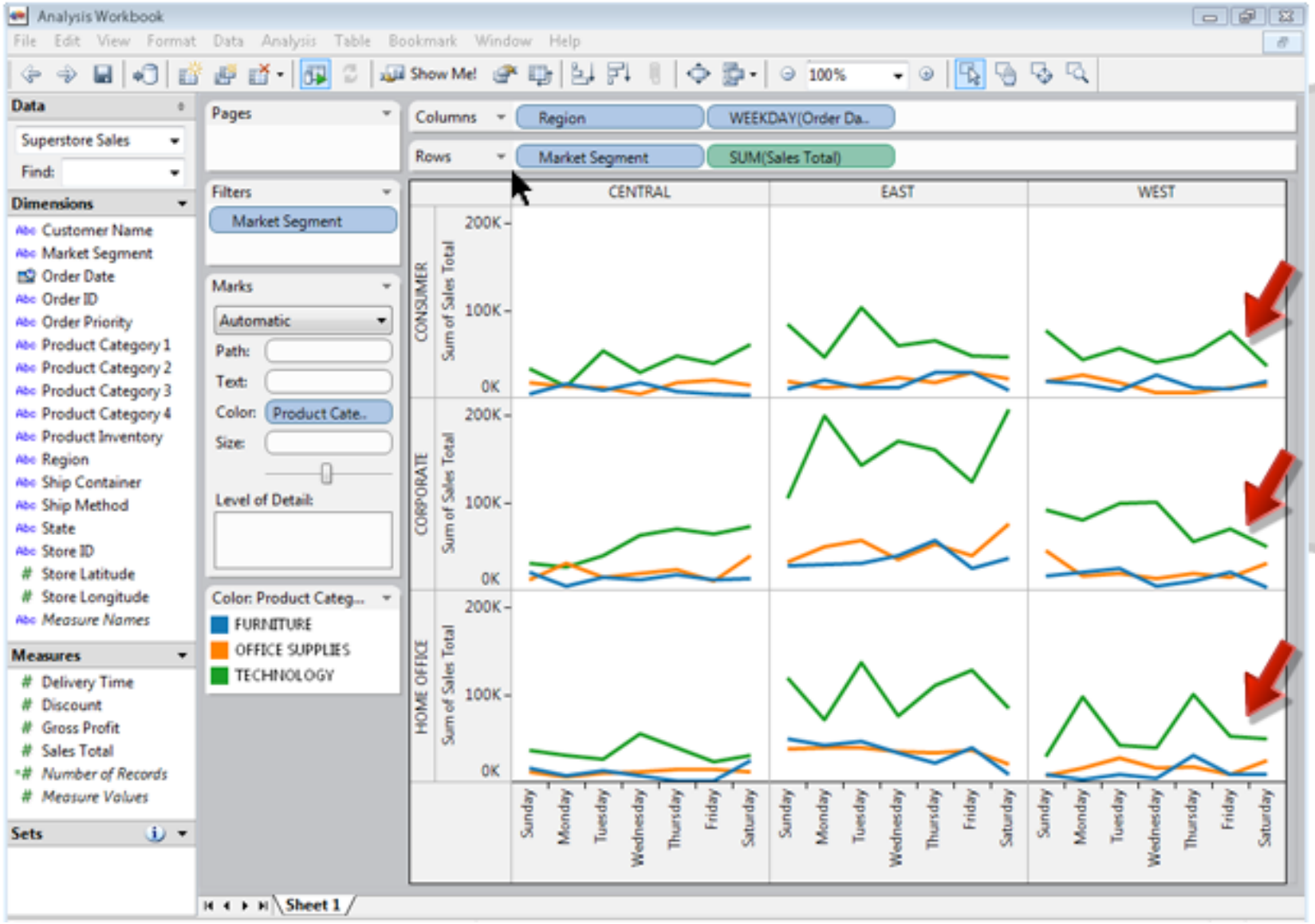
Multidimensional Data Model

- Recognized as a first class data model, with its own query language, UI, etc.
- Key to Executive Information Systems
  - widely used.
- No joins.
- Drill down for explanation.
- Usually read only, with heavy schema.
- Some direct manipulation.

UNDERSTAND TRENDS

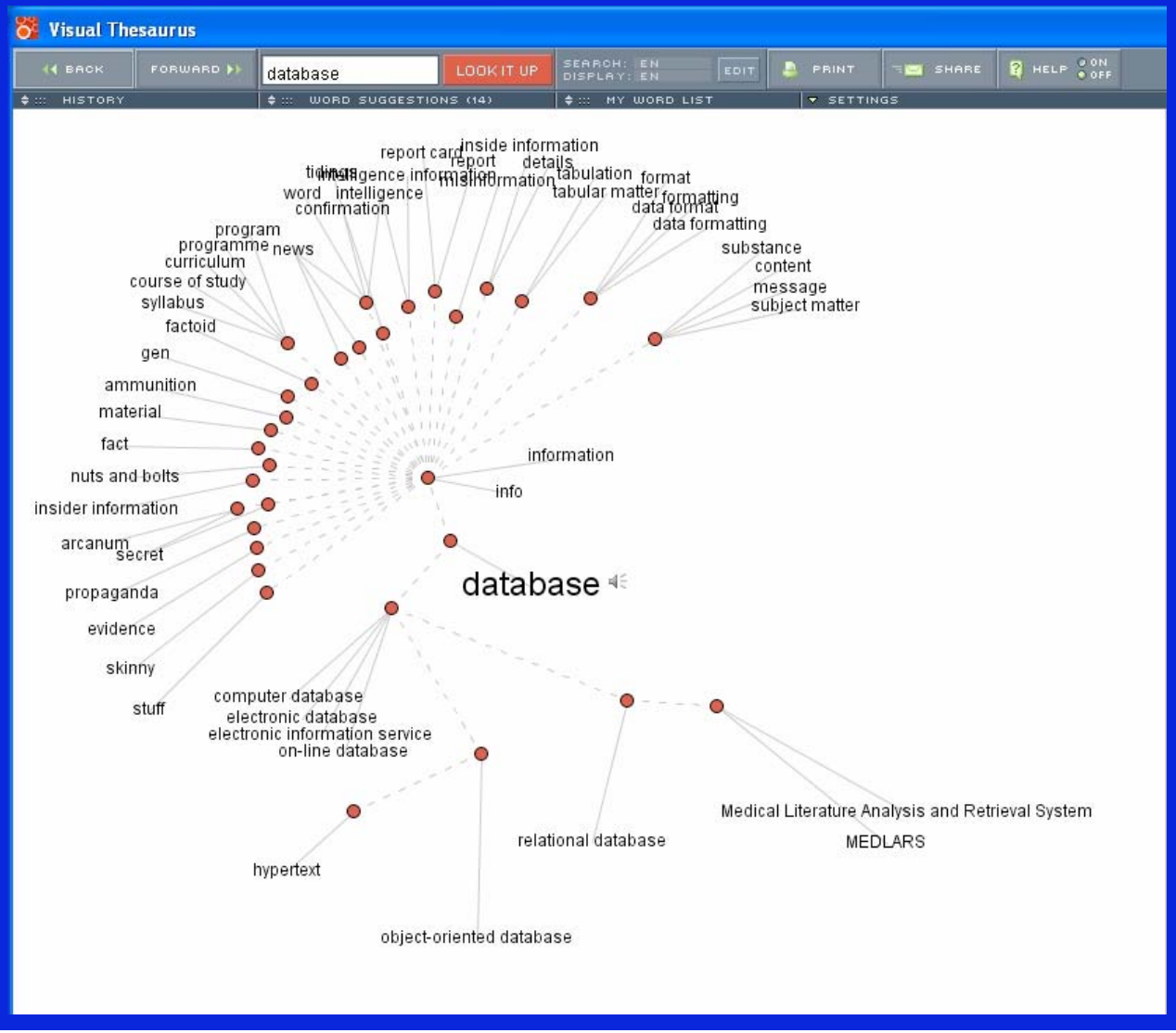
Analyze trends along multiple categories

TABLEAU SOFTWARE

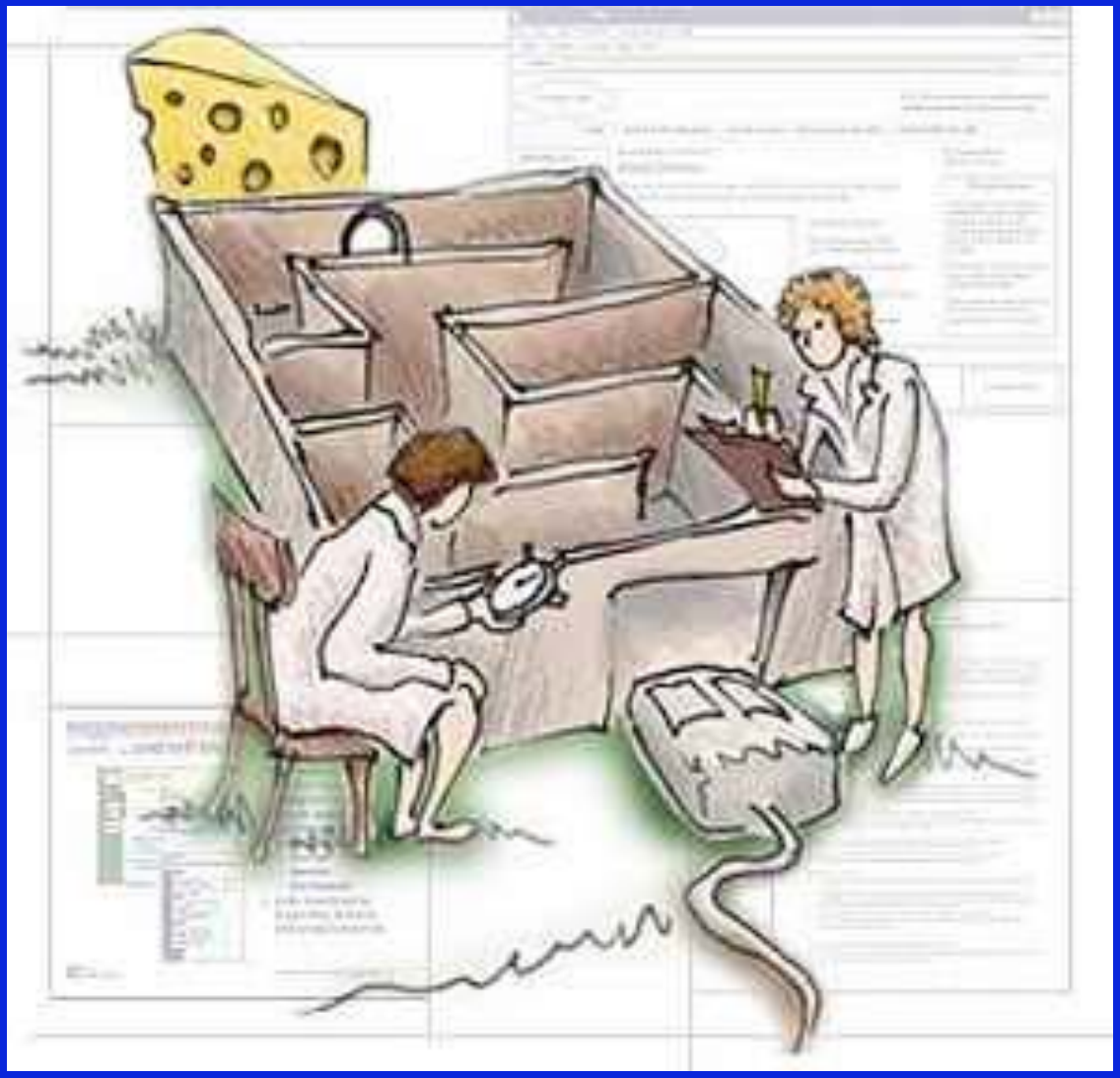


Database

Network Presentation Model



Traditional View of Usability



Usability Testing is Important

But …

Conclusion

- Biological data presents many interesting challenges that stress data management technology.
- Solutions to these challenges are likely to be of use in applications other than biological data management as well.
- We discussed some key aspects, including provenance, ontologies, and usability.

Bibliography

- Several references have been cited in context above. These are not repeated here.
- Given below are some additional relevant readings, grouped by topic.

Some Basic Readings

- H. Liu & L. Wong "Data mining tools for biological sequences", *JBCB*, 1:139-168, 2003
- J. Koh et al., "A Classification of Biological Data Artifacts", *DBiBD*, 2005
- *OMICS: A Journal of Integrative Biology*, Vol. 7, no.1, special issue on data management for biology, July 2003.
- *VLDB Journal*, Vol. 14, no. 3, special issue on data management, analysis, and mining for the life sciences, Sep. 2005.

Data Modeling Readings

- Data modeling
  - XML data modeling for relationships
    http://www.ibm.com/developerworks/xml/library/x-xdm2m.html
  - Data Modeling using XML Schemas – extremely detailed and very loooong tutorial.
    Murali Mani and Antonio Badia
    http://www.er.byu.edu/er2003/slides/ER2003PT2Mani.pdf
- GMOD
  - http://www.gmod.org/chado/
  - http://www.fruitfly.org/~cjm/chado-talk/chado-talk.html
  - Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, Lewis S. The generic genome browser: a building block for a model organism system database. Genome Res. 2002 Oct;12(10):1599-610. PMID: 12368253
    http://www.genome.org/cgi/reprint/12/10/1599.pdf

Data Modeling Readings (contd)

- GUS
  - http://www.gusdb.org/wiki/
  - http://www.gusdb.org/SchemaBrowser/
  - http://www.cbil.upenn.edu/~stoeckrt/ASM-GUS.ppt
  - Functional genomics databases on the web. Christian J. Stoeckert, Jr. Cellular Microbiology Volume 7 Issue 8 Page 1053 - August 2005. http://www.blackwell-synergy.com/doi/abs/10.1111/j.1462-5822.2005.00553.x

More Data Modeling Readings

- Gene expression data
  - A resource / repository
    http://www.ncbi.nlm.nih.gov/geo/
  - Microarray Gene Expression Data Society
    http://www.mged.org/
  - Minimum information for Microarray Experiments MIAME
    http://www.mged.org/Workgroups/MIAME/miame.html
  - MAGE Object Model
    http://www.omg.org/technology/documents/formal/gene_expression.htm
  - Graphical View (Rational Rose)
    http://www.ebi.ac.uk/arrayexpress-old/Schema/MAGE/MAGE.htm
  - DTD for MAGE
    http://xml.coverpages.org/MAGE-ML-dtd-2002-01-21.txt
  - Serial Analysis SAGE
    http://www.sagenet.org/findings/index.html
  - Detailed microarray and gene expression tutorials
    http://www.ims.nus.edu.sg/Programs/microarray/tutorial.htm

Data Integration Readings
Overview + Mediator solutions

- @article{ Ste03,
  Author = {Stein, L. D.},
  Title = {Integrating biological databases},
  Journal = {Nat Rev Genet}, Volume = {4}, Number = {5}, Pages = {337-345},
  Year = {2003} }
  http://www.umiacs.umd.edu/~louiqa/2006/828U/Protected/nrg1065.pdf

- @article{ HSK+01,
  Author = {Haas, Laura M. and Schwarz, P. M. and Kodali, P. and Kotlar, E. and Rice, J. and Swope, W. C.},
  Title = {DiscoveryLink: A System for Integrated Access to Life Sciences Data Sources},
  Journal = {IBM Systems Journal}, Volume = {40}, Number = {2}, Pages = {489-511}, Year = {2001} }
  http://www.research.ibm.com/journal/sj/402/haas.pdf

- @article{ ZLAE02,
  Author = {Zdobnov, Evgeni M. and Lopez, Rodrigo and Apweiler, Rolf and Etzold , Thure},
  Title = {The EBI SRS Server - Recent Developments},
  Journal = {Bioinformatics}, Volume = {18}, Number = {2}, Pages = {368-373},
  Year = {2002} }
  http://bioinformatics.oxfordjournals.org/cgi/reprint/18/2/368.pdf

Data Integration Readings
Mediation / Ontologies/ Warehouses

- @article{DCB+01,
  Author = {Davidson, Susan and Crabtree, Jonathan and Brunk, B.P. and Schug, Jonathan and Tannen, Val and Overton, G. Christian and Stoecker Jr., C. J .},
  Title = {K2/Kleisli and GUS: Experiments in Integrated Access to Genomic Data Sources},
  Journal = {IBM Systems Journal}, Volume = {40}, Number = {2}, Pages = {512-531}, Year = {2001} }
  http://www.research.ibm.com/journal/sj/402/davidson.pdf

- GeneExpress
  http://www.anthonykosky.com/anthol.html#gene_express

- @Article{Biowarehouse06,
  Author ="T.J. Lee and Y. Pouliot and V. Wagner and P. Gupta and D.W.J Stringer-Calvert and J.D. Tenenbaum and P.D. Karp",
  Title ="{BioWarehouse: a bioinformatics database warehouse toolkit}",
  journal ={BMC Bioinformatics}, volume ={7}, pages ={170}, year ={2006} }
  http://www.biomedcentral.com/content/pdf/1471-2105-7-170.pdf

Data Integration Readings
Entity Integrity + Semantics of answers

- Nucleic Acids Research 2005 January 1; 33(Database Issue):D54-D58; doi:10.1093/nar/gki031
  Entrez Gene: gene-centered information at NCBI.
  Donna Maglott*, Jim Ostell, Kim D. Pruitt, and Tatiana Tatusova
  http://nar.oxfordjournals.org/cgi/reprint/33/suppl_1/D54.pdf

- Nucleic Acids Research 2005 January 1; 33(Database Issue):D501-D504; doi:10.1093/nar/gki025
  NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.
  Kim D. Pruitt*, Tatiana Tatusova and Donna R. Maglott
  http://nar.oxfordjournals.org/cgi/reprint/33/suppl_1/D501.pdf

- Sarah Cohen-Boulakia, Susan Davidson, Christine Froidevaux
  A User-centric Framework for accessing Sources and Tools
  Proceedings of DILS'05, Data Integration in the Life Sciences,
  Springer-Verlag, LNCS series, Lecture Notes in Bioinformatics (LNBI), Vol. 3615, pp. 3-18, 2005.
  http://repository.upenn.edu/cgi/viewcontent.cgi?article=1241&context=cis_papers

Reading List on Provenance and Curation

- Peter Buneman, Adriane Chapman, James Cheney,
  "Provenance management in curated databases",
  in *Proceedings of the 2006 ACM SIGMOD international Conference on Management of Data* (Chicago, IL, USA, June 27-29, 2006),
  SIGMOD 2006, ACM Press, New York, NY, 539-550,
  http://portal.acm.org/citation.cfm?doid=1142473.1142534

- Yogesh L. Simmhan, Beth Plale, Dennis Gannon,
  "A survey of data provenance in e-science",
  SIGMOD Record, 34(3), September 2005, 31-36,
  http://portal.acm.org/citation.cfm?doid=1084805.1084812

- Chimera http://www.cgl.ucsf.edu/chimera/
  Ian Foster, Jens Vökler, Michael Wilde, Yong Zhao,
  "Chimera: a virtual data system for representing, querying, and automating data derivation",
  in *Proceedings of the 14th International Conference on Scientific and Statistical Database Management* (Edinburgh, Scotland, July 24-26, 2002),
  SSDBM 2002, 37-46,
  http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=1029704

More Provenance

- **ZOOM with user views**

  Shirley Cohen, Sarah Cohen Boulakia, Susan B. Davidson,
  "*Towards a Model of Provenance and User Views in Scientific Workflows*",
  in *the 3rd International Workshop on Data Integration in the Life Sciences 2006* (Hinxton, U.K., July 20-22, 2006),
  DILS 2006, Lecture Notes in Computer Science 4075, Springer, 264-279,
  http://www.springerlink.com/content/r123451r8104426u/

- **Provenance Challenge**

  http://twiki.ipaw.info/bin/view/Challenge/FirstProvenanceChallenge
  is a recent activity to provide a framework / dataset to compare the capabilities of systems that track provenance.

Usability Resource

- Usability is a new and open area
- Visit http://www.eecs.umich.edu/db/usable for more information