

Relational Databases for Biologists: Efficiently Managing and Manipulating Your Data

Session 1: Data Conceptualization and Database Design

George Bell, Ph.D.

WIBR Bioinformatics and Research Computing

What is a Database?

- A collection of data
- A set of rules to manipulate data
- A method to mold information into knowledge
- Is a phonebook a database?
 - Is a phonebook with a human user a database?

Babbitt, S.	38 William St., Cambridge	555-1212
Baggins, F.	109 Auburn Ct., Boston	555-1234
Bayford, A.	1154 William St., Newton	555-8934

Why are Databases Important?

- Data -> information -> knowledge
- Efficient manipulation of large data sets
- Integration of multiple data sources
- Adding crosslinks/references to other resources

Why is a Database Useful?

- If database systems simply manipulate data, why not use existing file system and spreadsheet mechanisms?
- “Baggins” Telephone No. Lookup:
 - Human: Look for B, then A, then G ...
 - Unix: `grep Baggins boston_directory.txt`
 - DB: `SELECT * FROM directory WHERE lastName="Baggins"`

Babbitt, S.	38 William St., Cambridge	555-1212
Baggins, F.	109 Auburn Ct., Boston	555-1234
Bayford, A.	1154 William St., Newton	555-8934

What is the Advantage of a Database?

- Find all last names that contain “th” but do not have street address that begin with “th”.
 - Human: a lot of careful reading....
 - Unix: Write a directory parser and a filter.
 - DB: `SELECT lastName FROM directory WHERE lastName LIKE "%th%" AND street NOT LIKE "Th%"`

Why Biological Databases?

- To access and manipulate lots of data
- To manage experimental results
- To improve search sensitivity
- To improve search efficiency
- To merge multiple data sets

Microarrays: a practical application

- The typical Excel spreadsheet of microarray data

Affy	lung	heart	gall_bladder	pancreas	testis
92632_at	20	20	20	20	20
94246_at	20	71	122	20	20
93645_at	216	249	152	179	226
98132_at	135	236	157	143	145

- Find all of the genes that have at least 2-fold higher expression in the gall bladder compared to the testis, and sort by decreasing RNA abundance in the heart

Course Goals

- Conceptualize data in terms of relations (database tables)
- Design relational databases
- Use SQL commands to extract data from (mine) databases
- Use SQL commands to build and modify databases

Session Outline

- Session 1
 - Database background and design
- Session 2
 - SQL to data mine a database
- Session 3
 - SQL to create and modify a database
- Hands-on sessions after each lecture

Supplemental Information

- Links to class information:
<http://jura.wi.mit.edu/bio/education/bioinfo2006/db4bio/>
- MySQL documentation:
<http://dev.mysql.com/doc/>
- Books:
 - *MySQL* – Paul DuBois
 - and many others

Flat vs. Relational Databases

- Flat file databases use identity tags or delimited formats to describe data and categories without relating data to each other
 - Most biological databases are flat files and require specific parsers and filters
- Relational databases store data in terms of their relationship to each other
 - A simple query language can extract information from any database

Fasta format sequence file

```
>gi|2137523|pir||I59068 MHC class I H2-K-b-alpha-2 cell surface glycoprotein - mouse (fragment)
AHTIQVISGCEVGSDGRLLRGYQQYAYDGC DYIALNEDLKTWTAADMAALITKHKWEQAGEAERLRAYLE
GTCVEWLRRYLKNGNATLLRT
```

```
>gi|25054197|ref|XP_193866.1| histocompatibility 2, K region [Mus musculus]
MSRGRGGWSRRGPSIGSGRHRKPRAMSRVSEWTLRTLLGYYNQSKGGSHTIQVISGCEVGSDGRLLRGYQ
QYAYDGC DYIALNEDLKTWTAADMAALITKHKWEQAGEAERLRAYLEGT CVEWLRRYLKNGNATLLRTDS
PKAHVTHHSRPEDKVT LRCWALGFYPADITLTWQLNGEELIQDMELVETRPAGDGT FQKWASVVVPLGKE
QYYTCHVYHQGLPEPLTLRWEPPPSTVSNMATVAVLVVLGAAIVTGAVVAFVMKMRRRNTGGKGGDYALA
PGSQTS DLSLPDCKVMVHDPHSLA
```

```
>gi|25032382|ref|XP_207061.1| similar to histocompatibility 2, K region [Mus musculus]
MVPCTLLLLLAAALAPTQTRAGPHSLRYFVTA VSRPGLGEP RYMEVG YVDDTEFVRF DSDAENPRYEPRA
RWMEQEGPEYWERETQKAKGNEQSFRVDLRTLLGYYNQSKGGSHTIQVISGCEVGSDGRLLRGYQQYAYD
GCDYIALNEDLKTWTAADMAALITKHKWEQAGEAERLRAYLEGT CVEWLRRYLKNGNATLLRTDSPKAHV
THHSRPEDKVT LRCWALGFYPADITLTWQLNGEELIQDMELVETRPAGDGT FQKWASVVVPLGKEQYYTC
HVYHQGLPEPLTLRWEPPPSTVSNMATVAVLVVLGAAIVTGAVVAFVMKMRRRNTGGKGGDYALAPGSQT
SDLSLPDCKVMVHDPHSLA
```

GenBank format sequence file

```
LOCUS       NM_205137                1111 bp    mRNA    linear   VRT 16-APR-2005
DEFINITION  Gallus gallus homeobox protein Nkx-2.8 (NKX2.8), mRNA.
ACCESSION   NM_205137 XM_444649
VERSION     NM_205137.1  GI:49170097
SOURCE      Gallus gallus (chicken)
  ORGANISM  Gallus gallus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Archosauria; Aves; Neognathae; Galliformes; Phasianidae;
            Phasianinae; Gallus.
FEATURES             Location/Qualifiers
     source            1..1111
                       /organism="Gallus gallus"
                       /mol_type="mRNA"
                       /db_xref="taxon:9031"
     gene              1..1111
                       /gene="NKX2.8"
                       /db_xref="GeneID:396037"
     CDS               32..613
                       /gene="NKX2.8"
                       /codon_start=1
                       /product="homeobox protein Nkx-2.8"
                       /protein_id="NP_990468.1"
                       /db_xref="GI:49170098"
                       /db_xref="GeneID:396037"
                       /translation="MLPTPFSVEDILSLEQSSAPGAPGVRRSPSVEEEPPSGQCLLSQ
PLQADQQQTDPCHHPKQPQRRKPRVLFSTQVLELERRFKQQKYLALEREHLANVLQ
LTSTQVKIWFQNRRYKCKRQRQDRSLEMATYPLPPRKVAVPVLRNGKPCFEGSQPHL
APYGITVSPYSYSTYYSAYGVSYGVTGVLTP"
ORIGIN
  1 cagggagctc acaccgatcc cccccggag gatgctgcc acccctttct cegtogagga
  61 taccctcagc ctggagcaga gcagcgctcc cggagccccc ggggtccgcc gcagccctc
```

The Relational Database

- Data is composed of sets of tables and links
- Structured Query Language (SQL) to query the database
- Database management system (DBMS) to manage the data

DBMS ACID properties

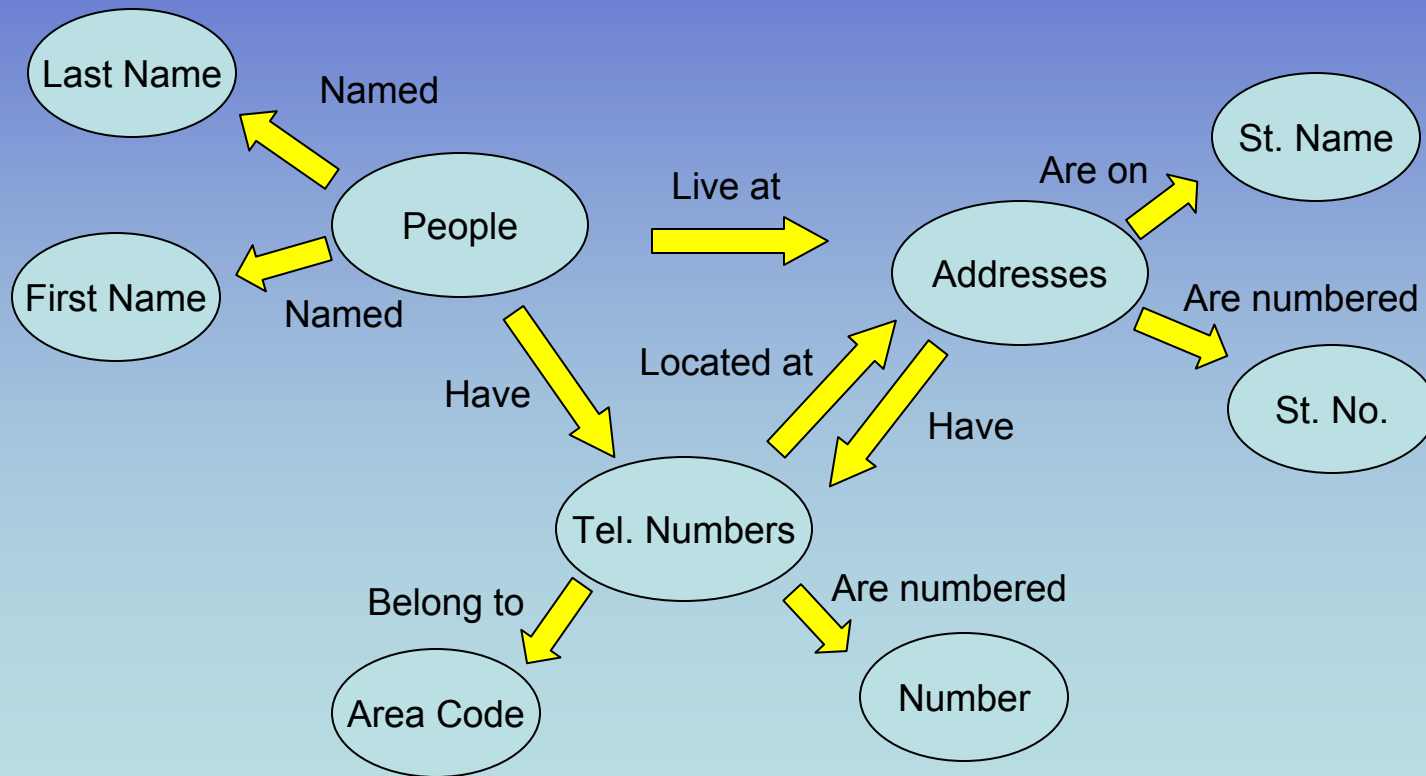
- ACID properties/model
 - Atomicity: “All or nothing” transactions
 - Consistency: Only valid data can be input
 - Isolation: Multiple user independence
 - Durability: Recovery mechanisms for system failures

Selected DBMSs

- MySQL
 - “The world's most popular open source database”, probably for biology too
 - Free; open source; small application; quick to learn
 - DBMS for this class
- PostgreSQL
 - “The world's most advanced open source database ”
 - Free; open source; somewhat larger application
- Oracle
 - “The worlds #1 database”
 - A lot more features but takes longer to learn
 - Expensive (but of course, many feel it’s worth it)
- All three are great choices and have the same core SQL functionality.

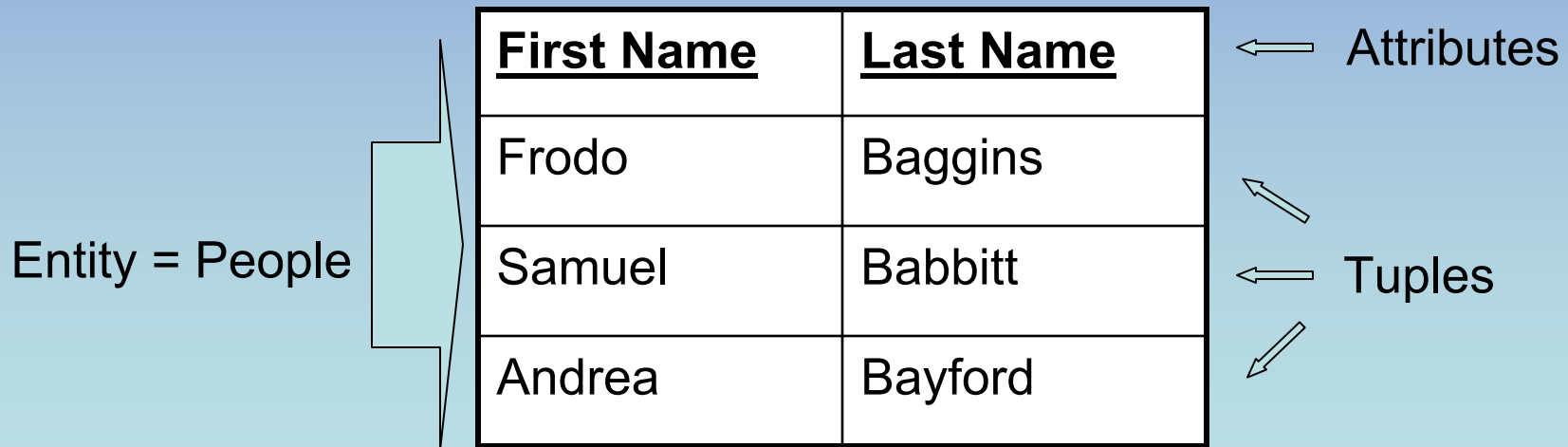
Data Conceptualization

- Data and Links (For a Phonebook)



Data Structure

- Data stored in tables with multiple columns (“attributes”).
- Each record is represented by a row (a “tuple”)



Relational Database Specifics

- Tables are relations
 - You perform operations on the tables
- No two tuples (rows) should be identical
- Each attribute for a tuple has only one value
- Tuples within a table are unordered
- Each tuple is uniquely Identified by a primary key

Primary Keys

- Primary Identifiers (IDs)
- Set of attributes that uniquely define a single, specific tuple (row)
- Must be absolutely unique
 - SSN ?
 - Phone Number ?
 - ISBN ?

First Name	Last Name	<u>SSN</u>
Frodo	Baggins	332-97-0123
Frodo	Binks	398-76-5327
Maro	Baggins	215-01-3965

Find the Keys

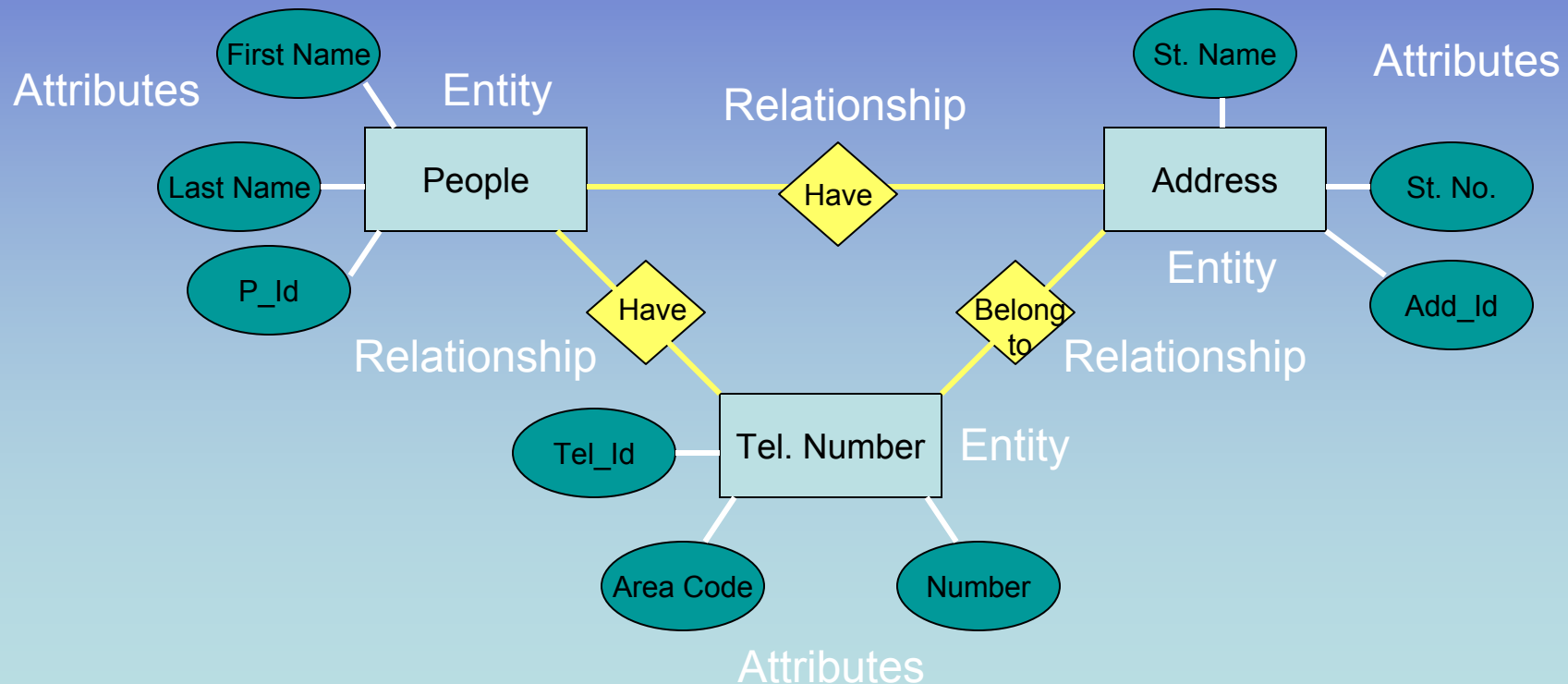
First Name	Last Name	SSN	Phone Number	Address
Frodo	Baggins	321-45-7891	123-4567	29 Hobbitville
Aragon	Elf-Wantabe	215-87-7458	258-6109	105 Imladris
Boromir	Ringer	105-91-0124	424-9706	31 Hobbitville
Bilbo	Baggins	198-02-2144	424-9706	29 Hobbitville
Legolas	Elf	330-78-4230	555-1234	135 Imladris

Design Principles

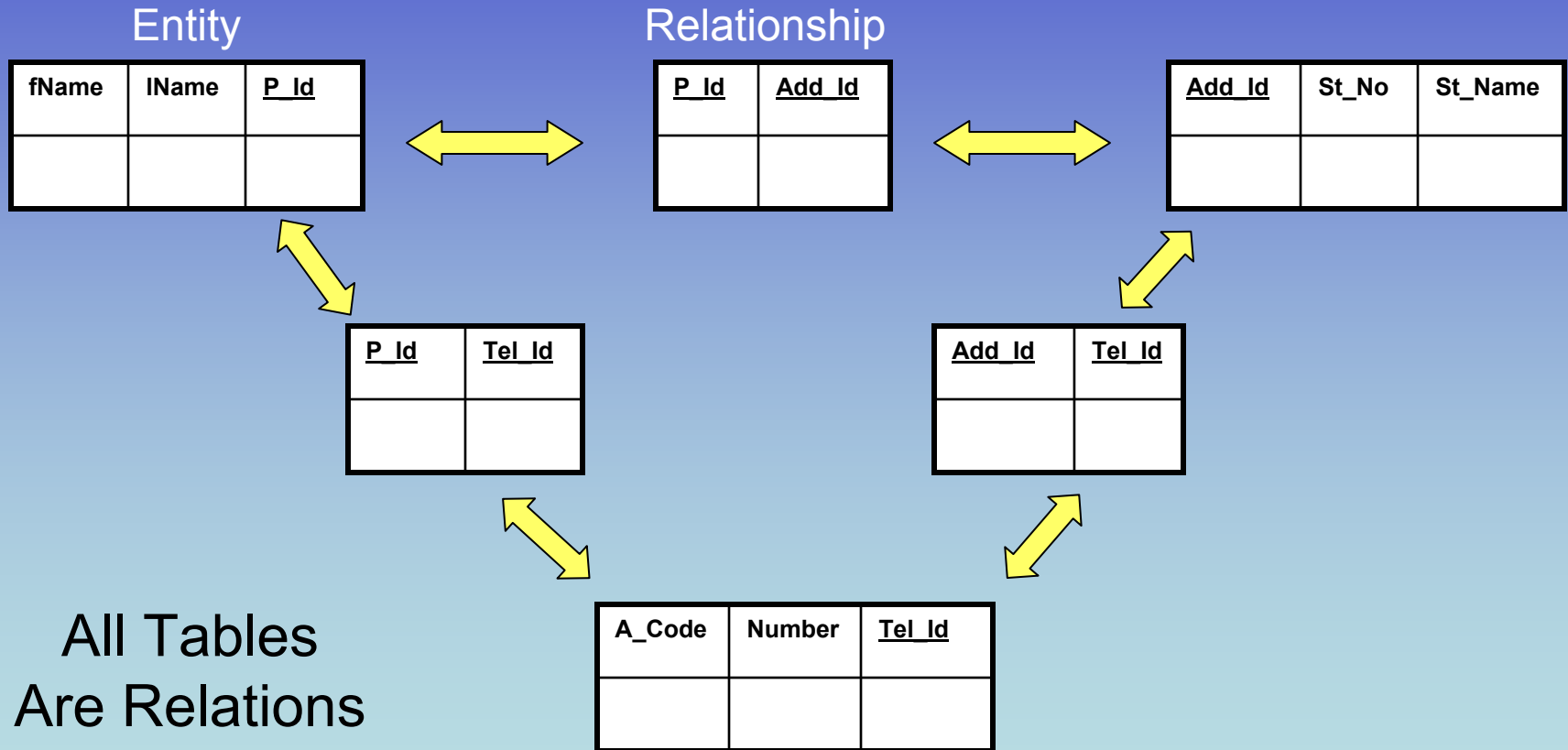
- Conceptualize the data elements (entities)
- Identify how the data is related
- Make it simple
- Avoid redundancy
- Make sure the design accurately describes the data!

Entity-Relationship Diagrams

- Expression of a database table design



E-R to Table Conversion



Steps to Build an E-R Diagram

- Identify data attributes
- Conceptualize entities by grouping related attributes
- Identify relationships/links
- Draw preliminary E-R diagram
- Add cardinalities and references

Developing an E-R Diagram

- Convert a GenBank File into an E-R Diagram

```
LOCUS       NM_205137                1111 bp    mRNA     linear   VRT 16-APR-2005
DEFINITION  Gallus gallus homeobox protein Nkx-2.8 (NKX2.8), mRNA.
ACCESSION  NM_205137 XM_444649
VERSION    NM_205137.1  GI:49170097
SOURCE     Gallus gallus (chicken)
ORGANISM   Gallus gallus
           Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
           Archosauria; Aves; Neognathae; Galliformes; Phasianidae;
           Phasianinae; Gallus.
FEATURES   Location/Qualifiers
    source  1..1111
            /organism="Gallus gallus"
            /mol_type="mRNA"
            /db_xref="taxon:9031"
    gene    1..1111
            /gene="NKX2.8"
            /db_xref="GeneID:396037"
    CDS     32..613
            /gene="NKX2.8"
            /codon_start=1
            /product="homeobox protein Nkx-2.8"
            /protein_id="NP_990468.1"
            /db_xref="GI:49170098"
            /db_xref="GeneID:396037"
            /translation="MLPTPFSVEDILSLEQSSAPGAPGVRRSPSVEEEEPPSGQCLLSQ
            PLQADQQQTDPCCHPKQPQRRKPRVLFSSQTQVLELERRFKQOKYLSALEREHLANVLQ
            LTSTQVKIWFQNRRYKCKRQRQDRSLEMATYPLPPRKVAVPVLVVRNGKPCFEFSQPHL
            APYGITVSPYSYSTYYSAYGVSYGVTGVLTP"
ORIGIN
1  caggagacct caccagatcc ccccaggag gatctgtccc accctttct cggtaggagga
61  tctctcagc ctggagcaga gcagcgctcc cggagcccc ggggtccgcc gcagcccttc
```

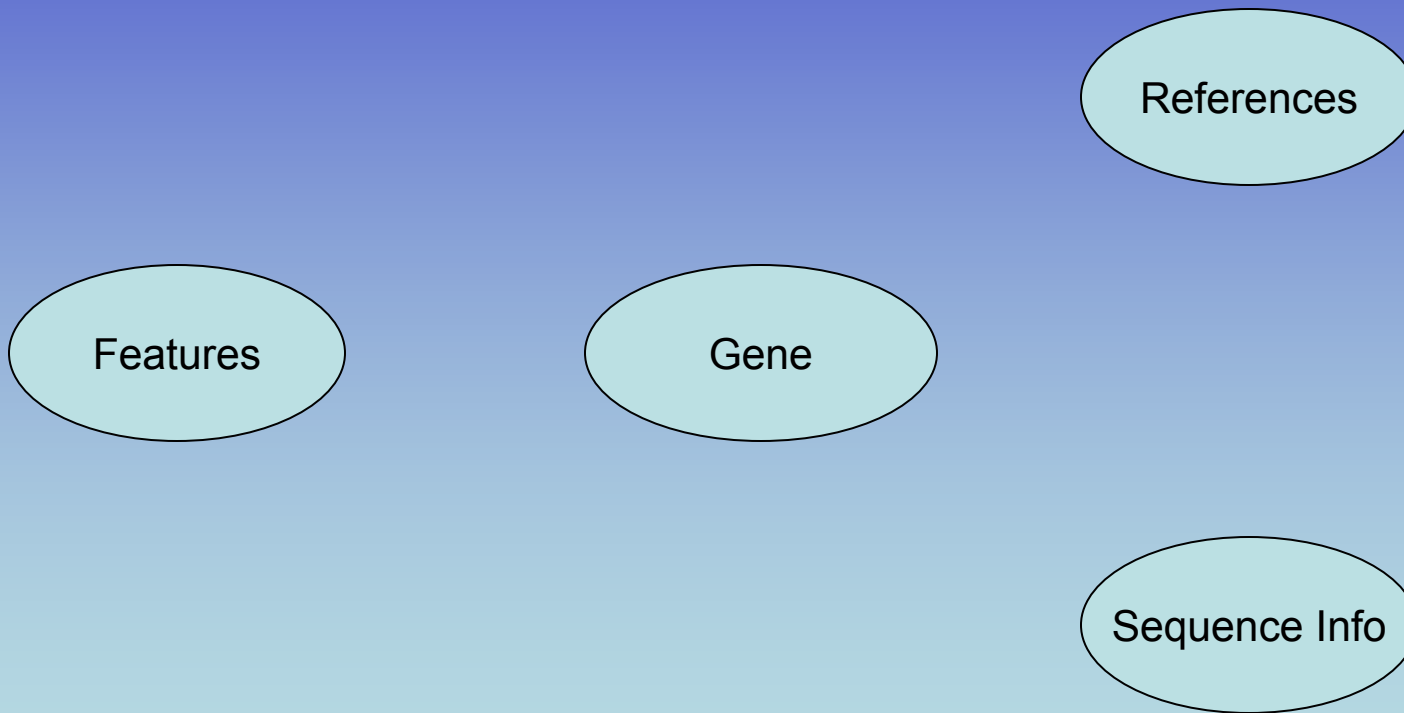
Identify Attributes

- Locus, Definition, Accession, Version, Source Organism
- Authors, Title, Journal, Medline Id, PubMed Id
- Protein Name, Protein Description, Protein Id, Protein Translation, Locus Id, GI
- A count, C count, G count, T count, Sequence

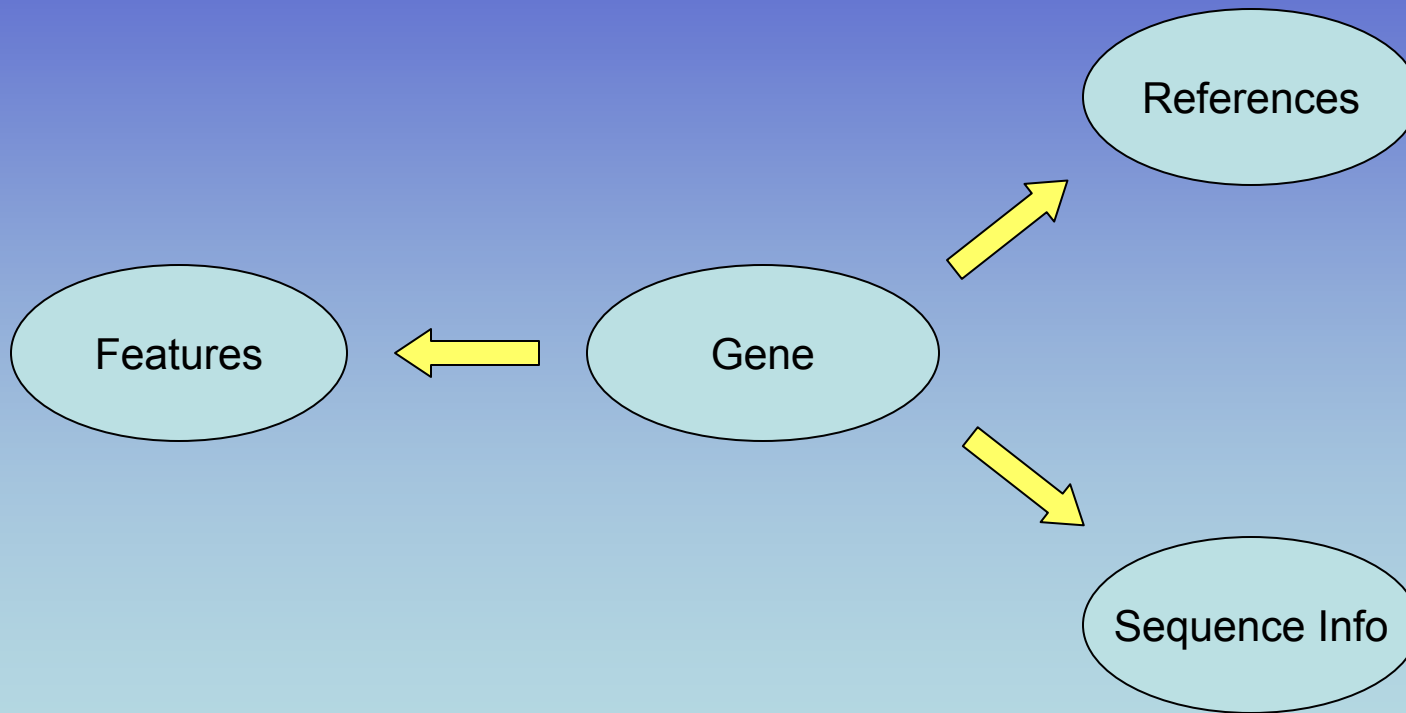
Identify Entities by Grouping

- Gene
 - Locus, Definition, Accession, Version, Source Organism
- References
 - Authors, Title, Journal, Medline Id, PubMed Id
- Features
 - Protein Name, Protein Description, Protein Id, Protein Translation, Locus Id, GI
- Sequence Information
 - A count, C count, G count, T count, Sequence

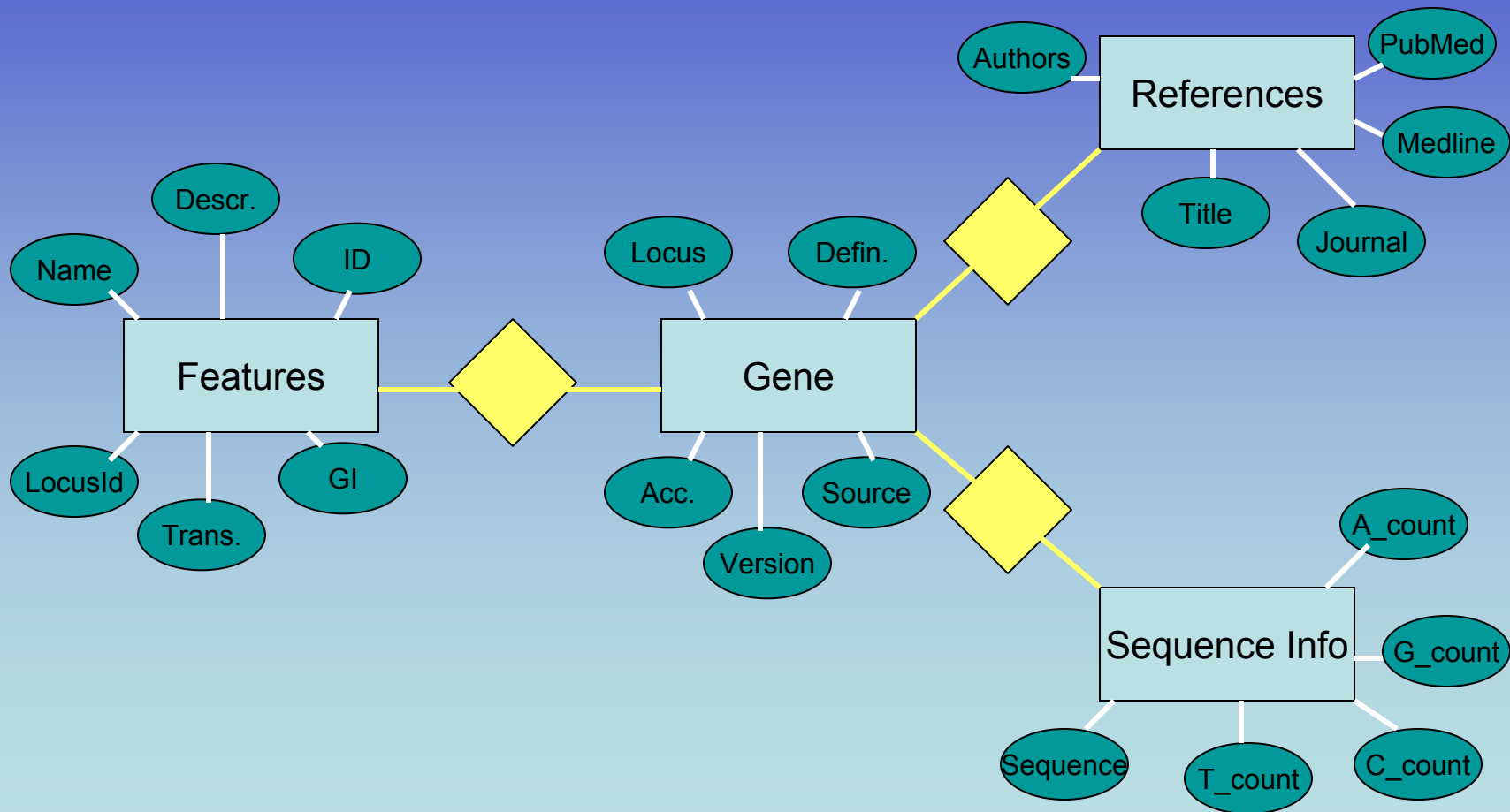
Conceptualize Entities



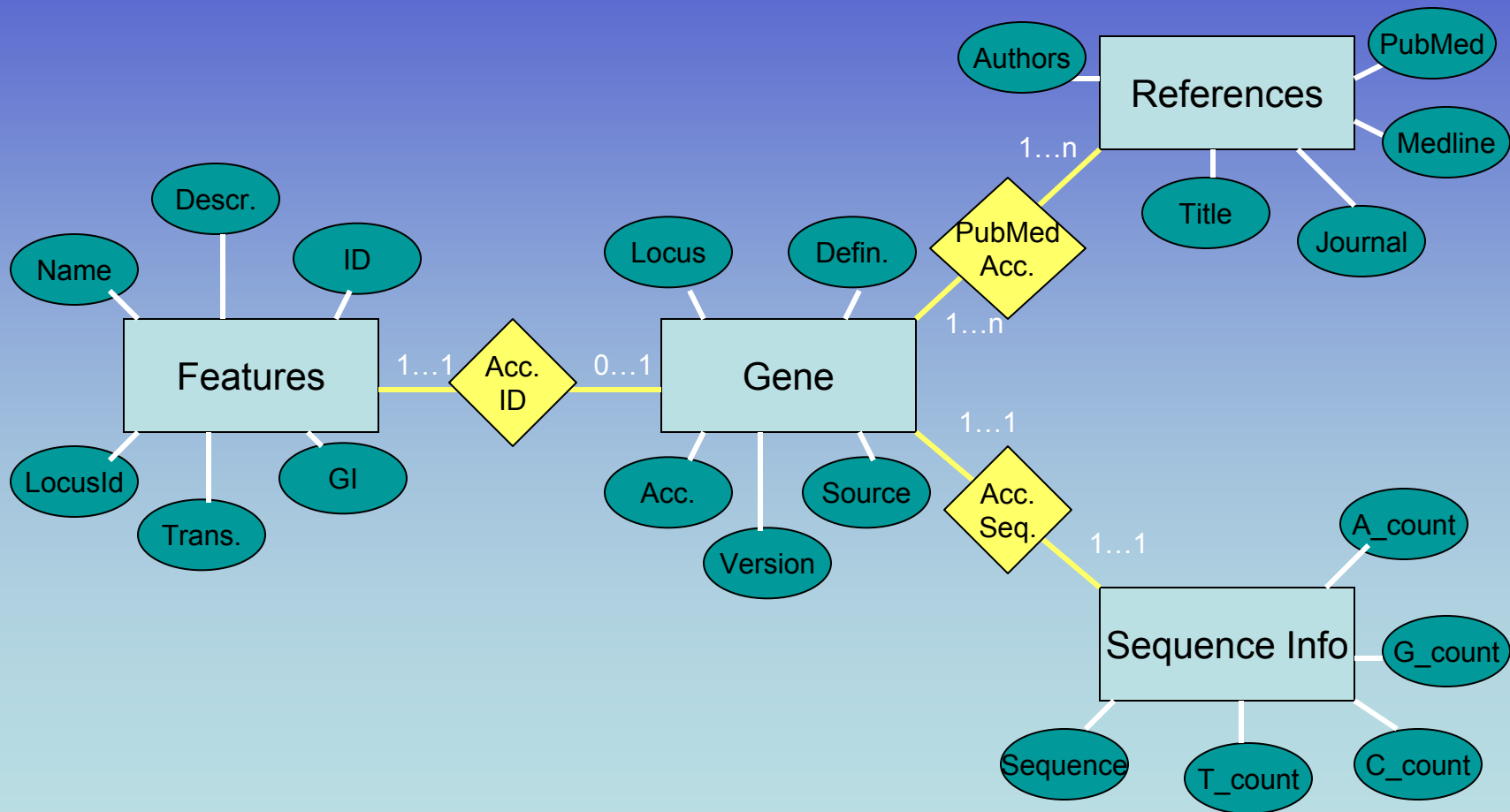
Identify Relationships



Preliminary E-R Diagram

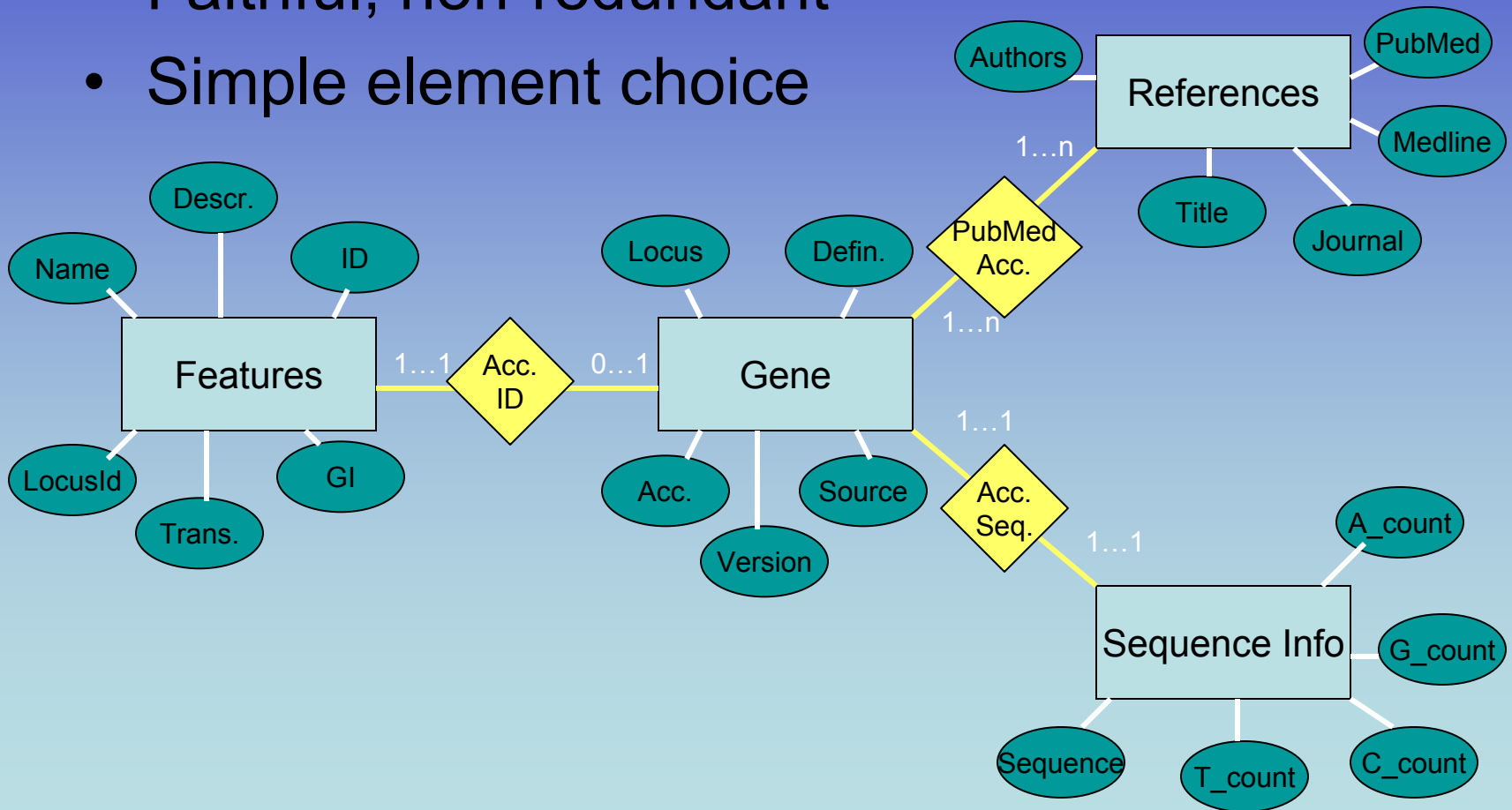


Cardinalities and References



Apply Design Principles

- Faithful, non-redundant
- Simple element choice



Summary

- Databases provide ACID
- Databases are composed of tables (relations)
- Relations are entities that have attributes and tuples
- Databases can be designed from E-R diagrams that are easily converted to tables
- Primary keys uniquely identify individual tuples and represent links between tables

Next class

- Using structured query language (SQL) to data mine databases
- `SELECT a FROM b WHERE c = d`

Database design example:

Design the db4bio database

Build Your Own E-R Diagram

- Express the following annotated microarray data set as an E-R diagram

<u>Affyld</u>	<u>GenBankId</u>	<u>Name</u>	<u>Description</u>	<u>LocusLinkId</u>	<u>LocusDescr</u>	<u>NT</u>	<u>RefSeq AA</u>	<u>RefSeq</u>	<u>\\</u>
U95-32123_at	L02870	COL7A1	Collagen	1294	Collagen	NM_000094	NP_000085	\\	
U98-40474_at	S75295	GBE1	Glucan	2632	Glucan	NM_000158	NP_000149	\\	

<u>Unigeneld</u>	<u>GO Acc.</u>	<u>GO Descr.</u>	<u>Species</u>	<u>Source</u>	<u>Level</u>	<u>Experiment</u>
Hs.1640	0005202	Serine Prot.	Hs	Pancreas	128	1
Hs.1691	0003844	Glucan Enz.	Hs	Liver	57	2

Identify Attributes

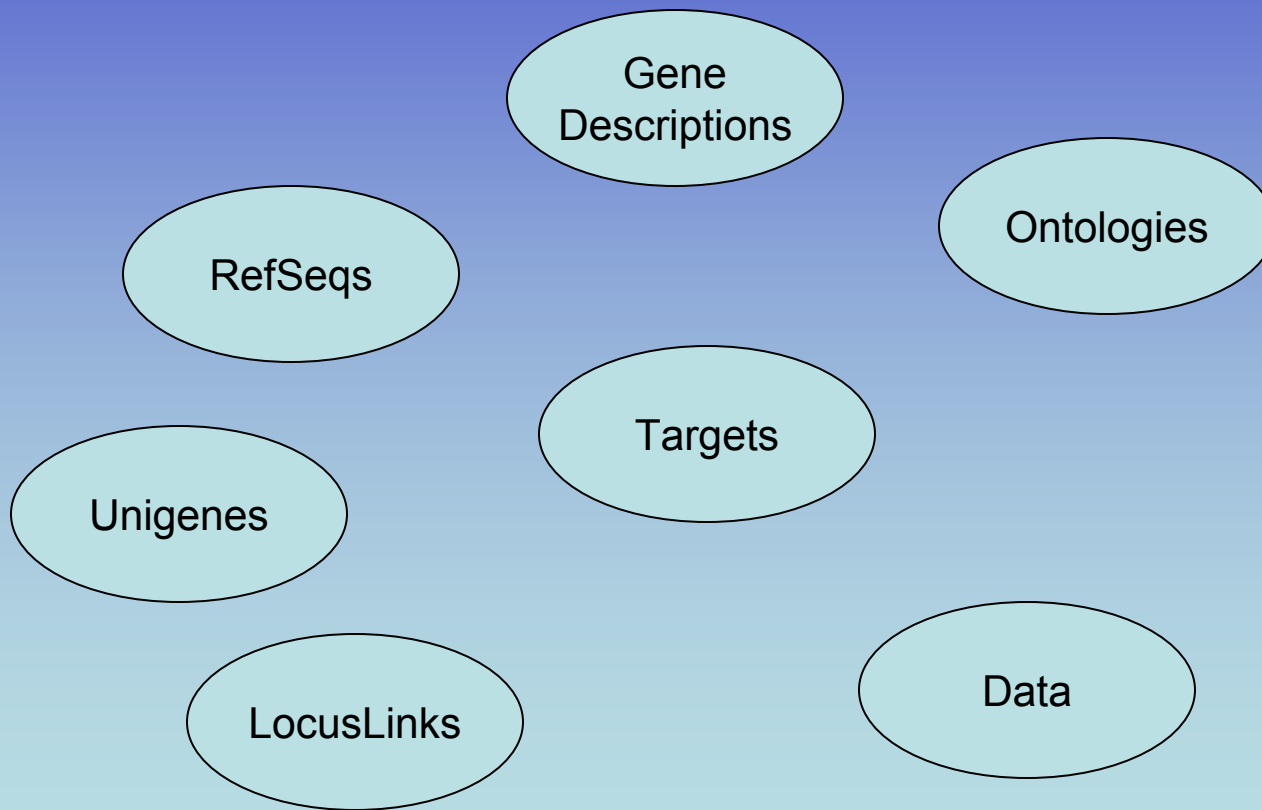
<u>AffyId</u>	<u>GenBankId</u>	<u>Name</u>	<u>Description</u>	<u>LocusLinkId</u>	<u>LocusDescr</u>	<u>NT</u>	<u>RefSeq</u>	<u>AA</u>	<u>RefSeq</u>	<u>\\</u>
U95-32123_at	L02870	COL7A1	Collagen	1294	Collagen	NM_000094	NP_000085			\\
U98-40474_at	S75295	GBE1	Glucan	2632	Glucan	NM_000158	NP_000149			\\

<u>UnigenelId</u>	<u>GO Acc.</u>	<u>GO Descr.</u>	<u>Species</u>	<u>Source</u>	<u>Level</u>	<u>Experiment</u>
Hs.1640	0005202	Serine Prot.	Hs	Pancreas	128	1
Hs.1691	0003844	Glucan Enz.	Hs	Liver	57	2

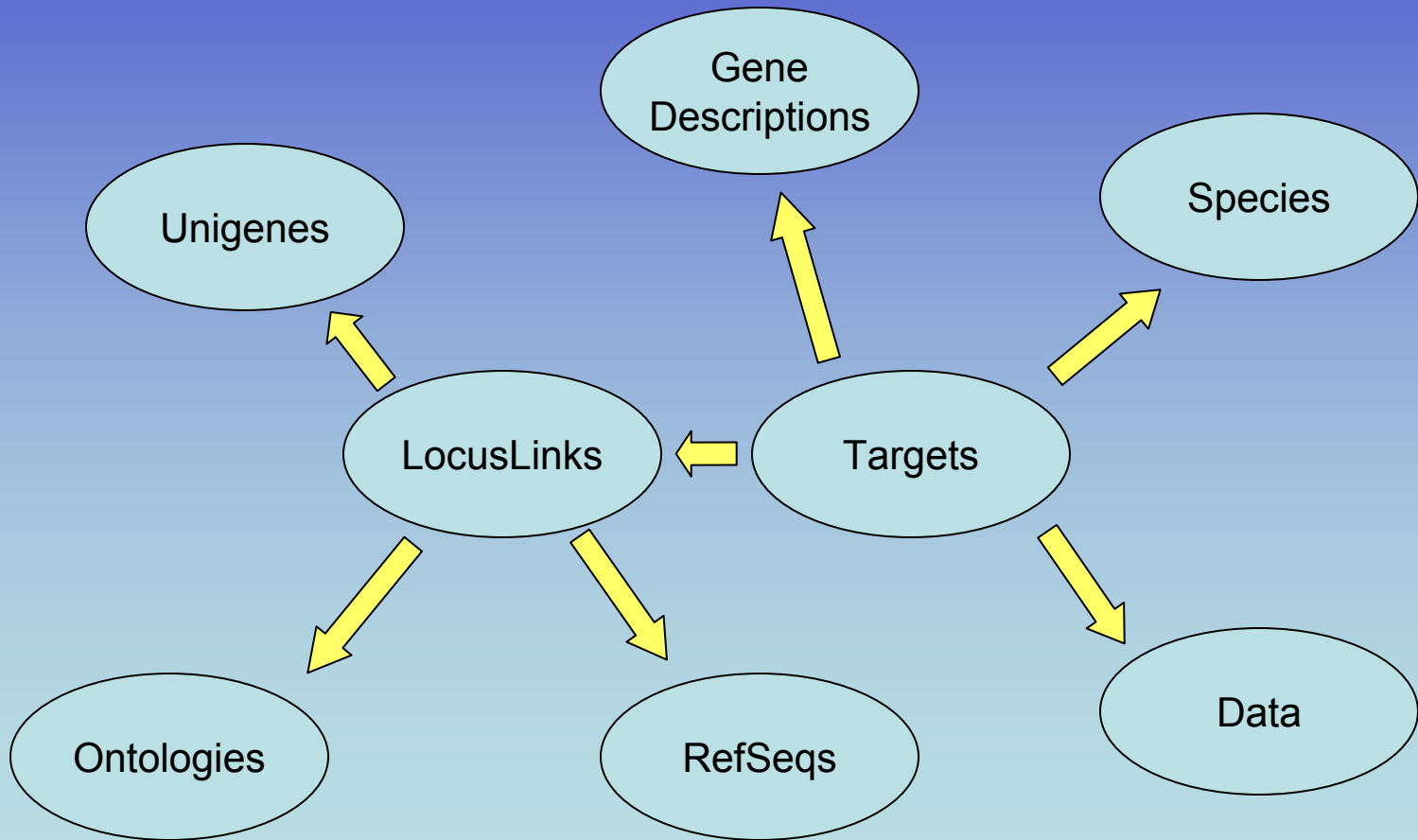
Identify Entities by Grouping

- Gene Descriptions
 - Name, Description, GenBank
- RefSeqs
 - NT RefSeq, AA RefSeq
- Ontologies
 - GO Accession, GO Terms
- LocusLinks
- Unigenes
- Data
 - Sample Source, Level
- Targets
 - Affy ID, Experiment Number, Species

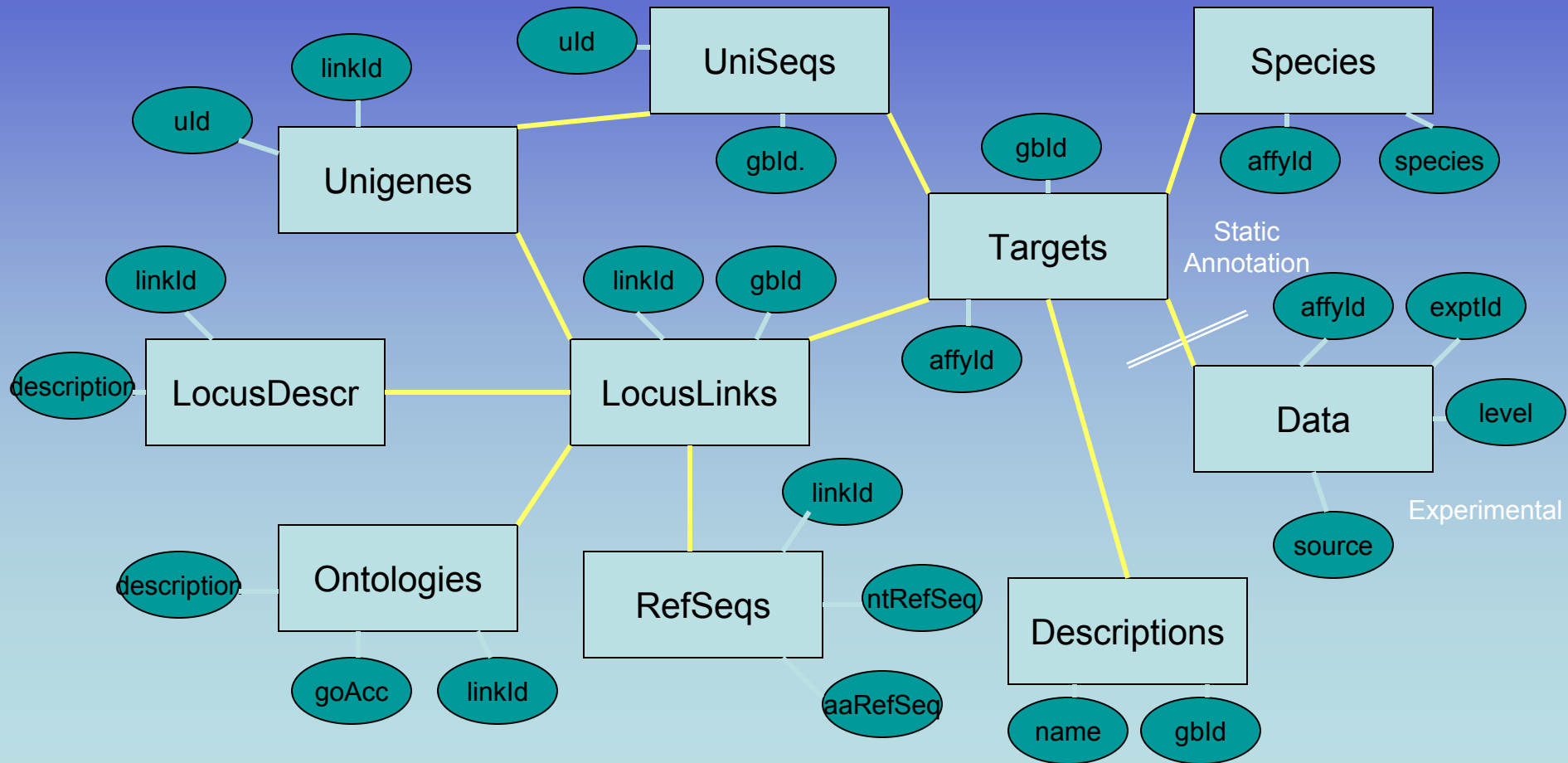
Conceptualize Entities



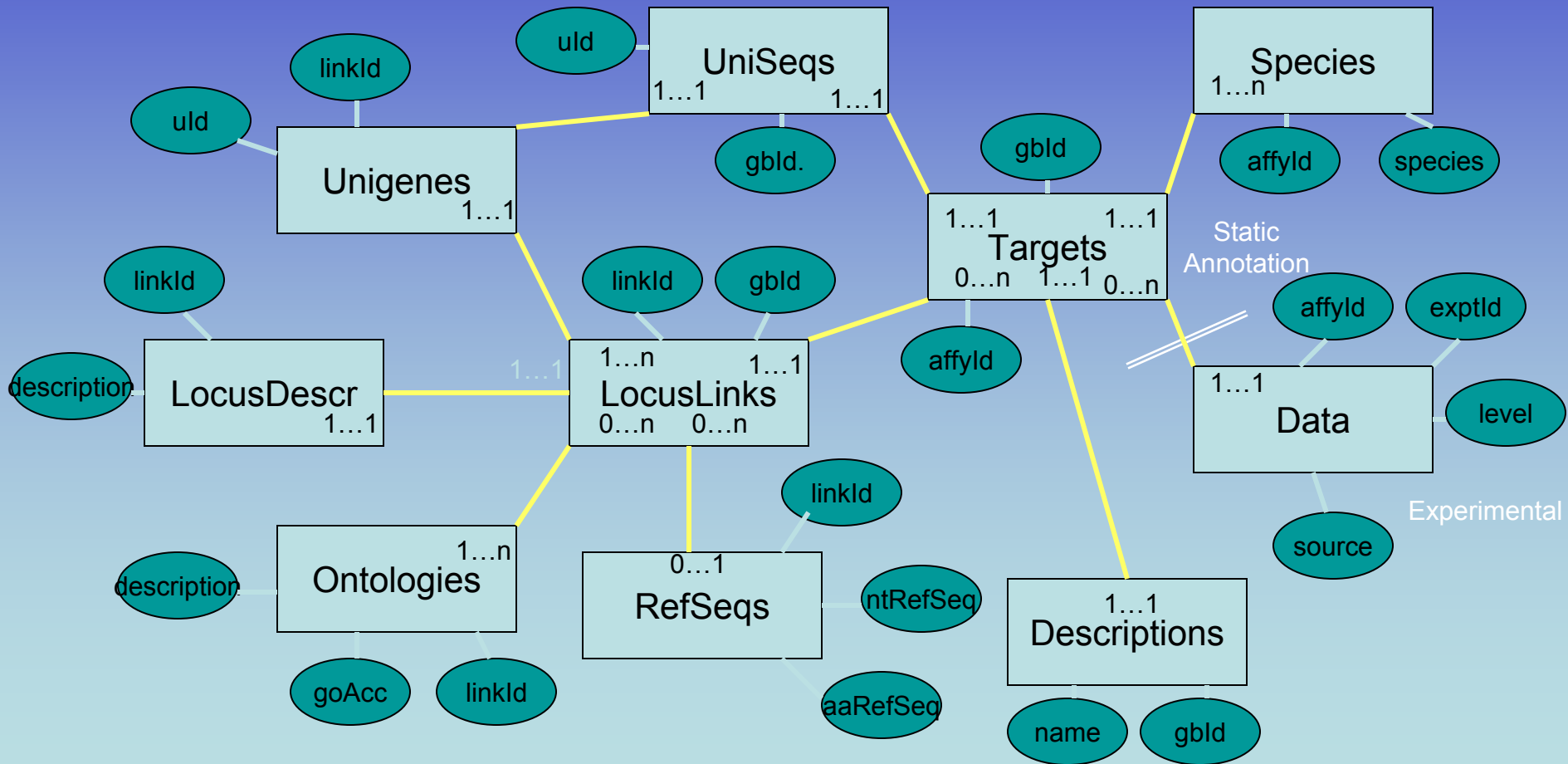
Identify Relationships



Preliminary E-R Diagram



Cardinalities and References



Apply Design Principles

