BINF 8211/6211
Design and Implementation of Bioinformatics Databases
Lecture #4

Dr. D. Andrew Carr
Dept. Bioinformatics and Genomics UNCC
Spring 2016

Class Information

- Class website
- http://ponden.github.io/databases2016
- User: student :: $tUd3Nt
- Homework 1
  - Due 8:00 a.m. today

Question:

- What are the 4 levels of data model?

- Relational Model Characteristics?

Data Model Levels

- External
  - End user view
  - Basic representation
- Conceptual.
  - Linkage to schema
  - Greater detail
- Internal model
  - Code, Script, Implementation
- Physical Model
  - How the data is actually stored
    - Binary

Relational Table Characteristics

- Two dimensional (rows and columns)
- Each row (tuple) is a single entity set.
- Every column has distinct name (attribute)
- Row+column = single data value
- All values in the same format within a column
- Columns have specific ranges (attribute domain)
- Order of rows and columns is permutable
- There exists at least one attribute(s) that uniquely identifies each row

ER – Model cont'd

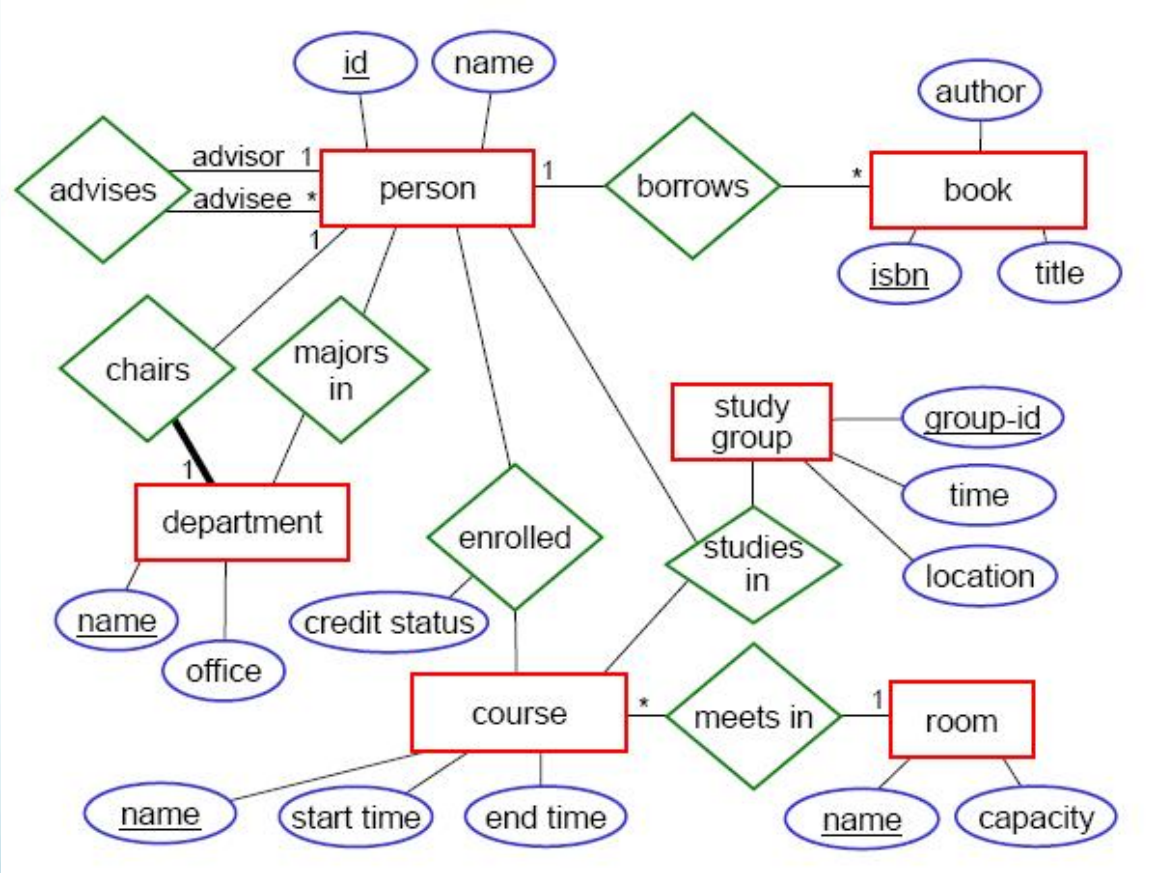


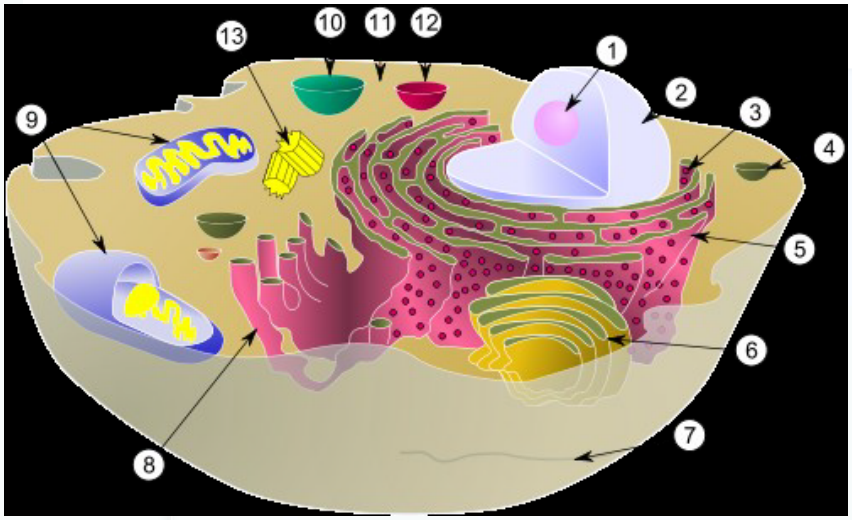
Image Taken from Web:http://www.snipview.com/q/Entity%E2%80%93relationship_model

Biological Example

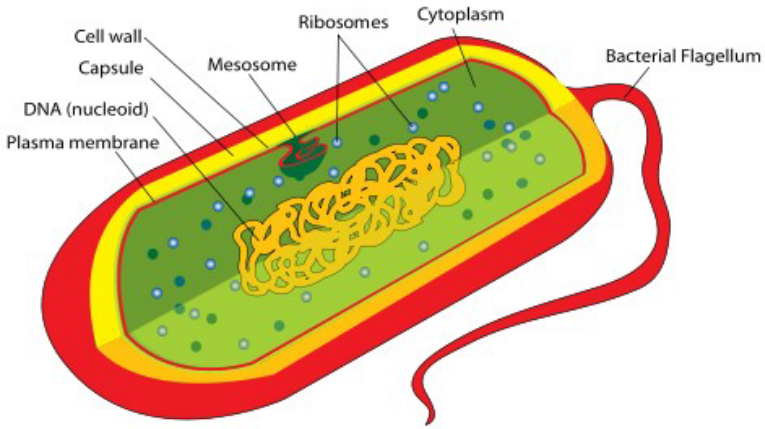
- To illustrate the concepts, vocabulary and approach we will examine am a use case in which we want to predict the presence of bacterial operons by determining whether genes that are close together also have expression levels that are very similar under nearly all conditions.

Example from J. Weller 2015

Conceptual biological representations (models) of cells.



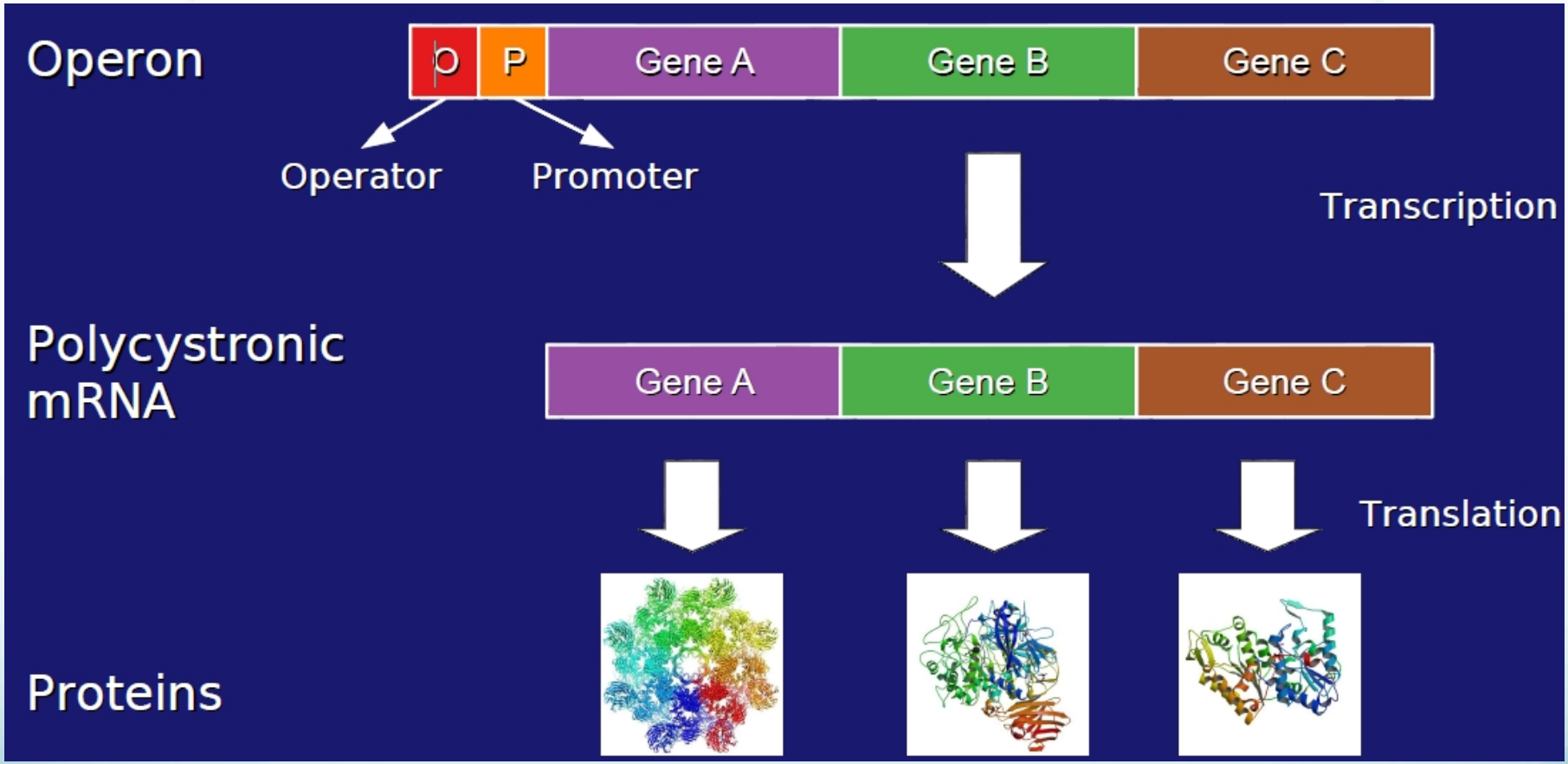
Type: Eukaryote
Diameter: 10-100um
HAS Nucleus
HAS Organelles
IS Unicellular OR multicellular



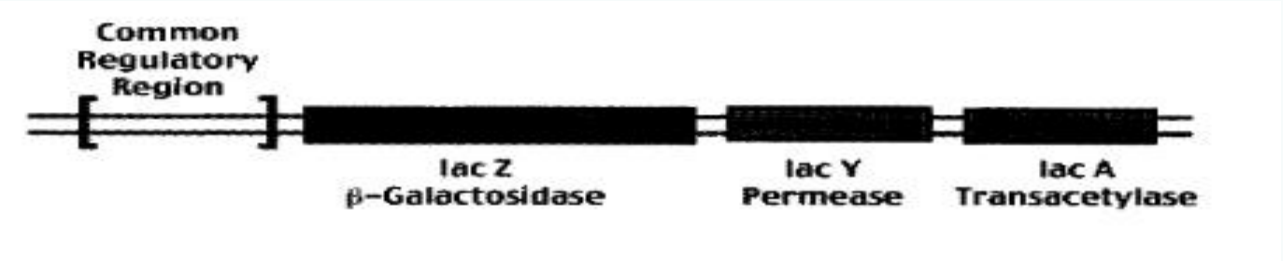
Type: Prokaryote
Diameter: 1-2um
HAS Nucleoid
HAS NO Organelles
IS Unicellular

Rules: A cell cannot be both a eukaryote and a prokaryote.
A cell must be one or the other.

Another model: mRNA in prokaryotes



The 3 proteins are translated at the same time.

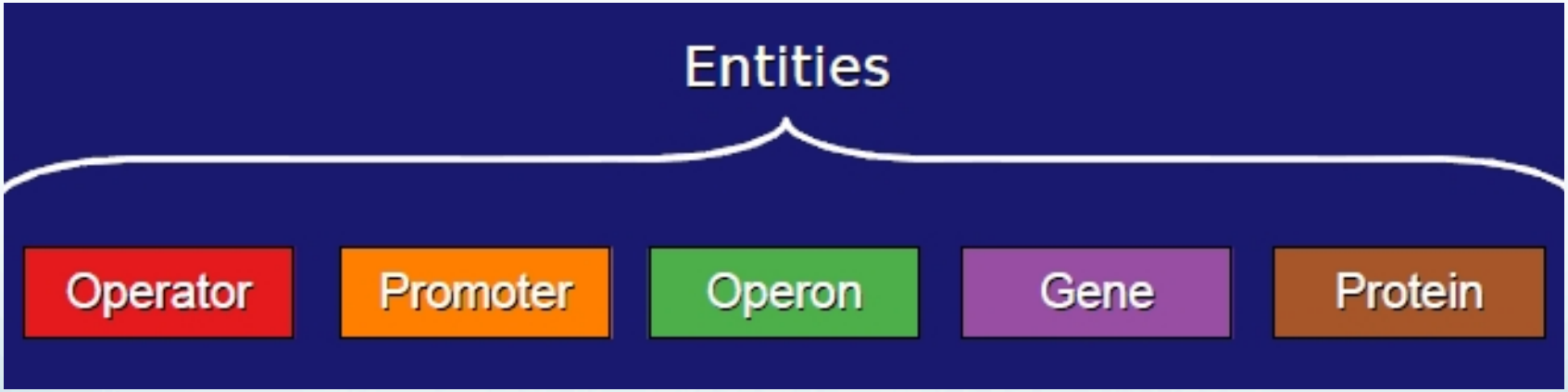


- Operon
  - HAS-A Regulatory region
    - HAS operator
    - HAS promotor
  - HAS gene (≥ 3?)
    - Has polycistronic transcript
    - HAS coding sequence
  - HAS ribosome binding site
  - HAS a terminator (≥ 1?)
  - Process: Transcription initiation Machinery IS-A complex of [sigma + polymerase]
  - Process: Operator binds activator/repressor
  - Process: Repressor Mechanism IS steric hindrance

Defined Relationships

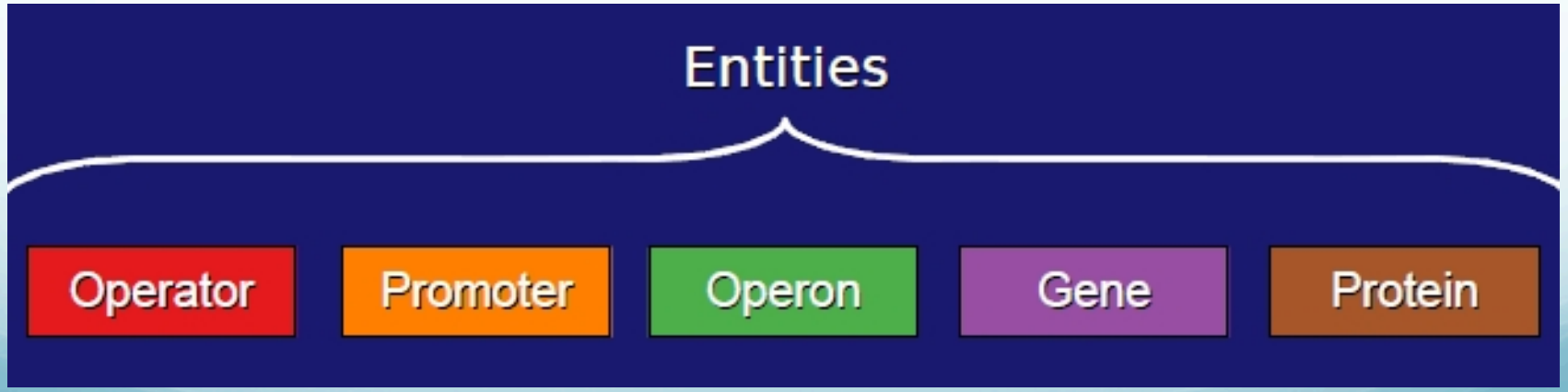
Operon Model

- Using operon biological construct, what are the entities?



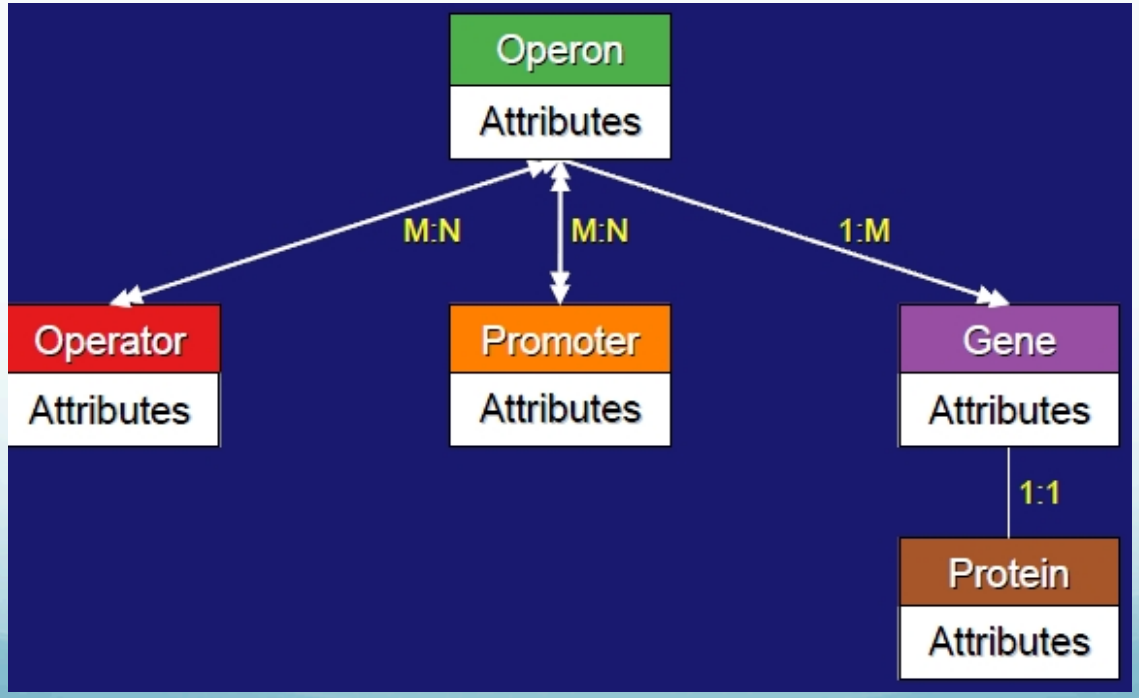
Operon Model

- An entity is a thing you can name (a noun) that has properties whose descriptions you want to store.
- Nouns: person, place, object, event, idea
- Going to our operon biological construct, what are the entities?



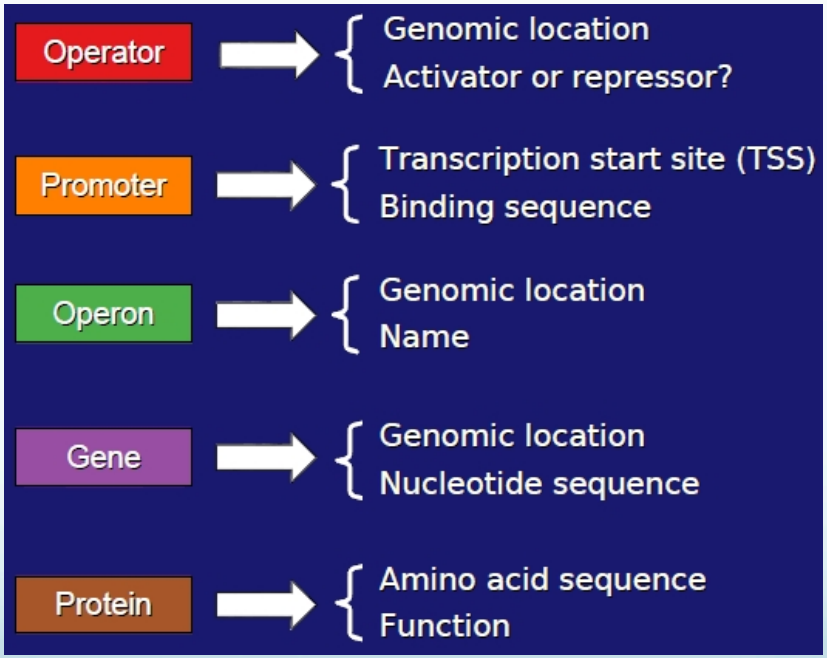
Operon Model

- An Entity Relationship Diagram gives a visual representation of the elements as you have defined them – laying them out this way let's you spot problems and share the model with others.

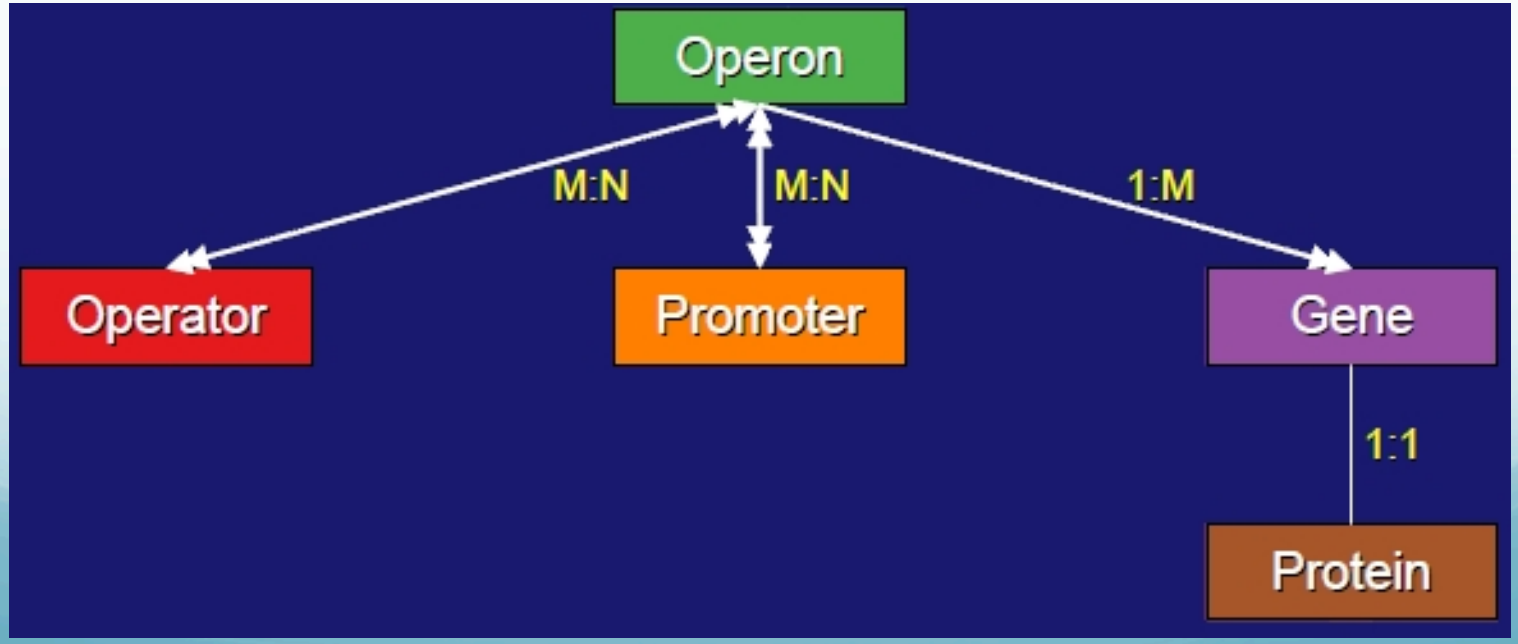


Operon Model

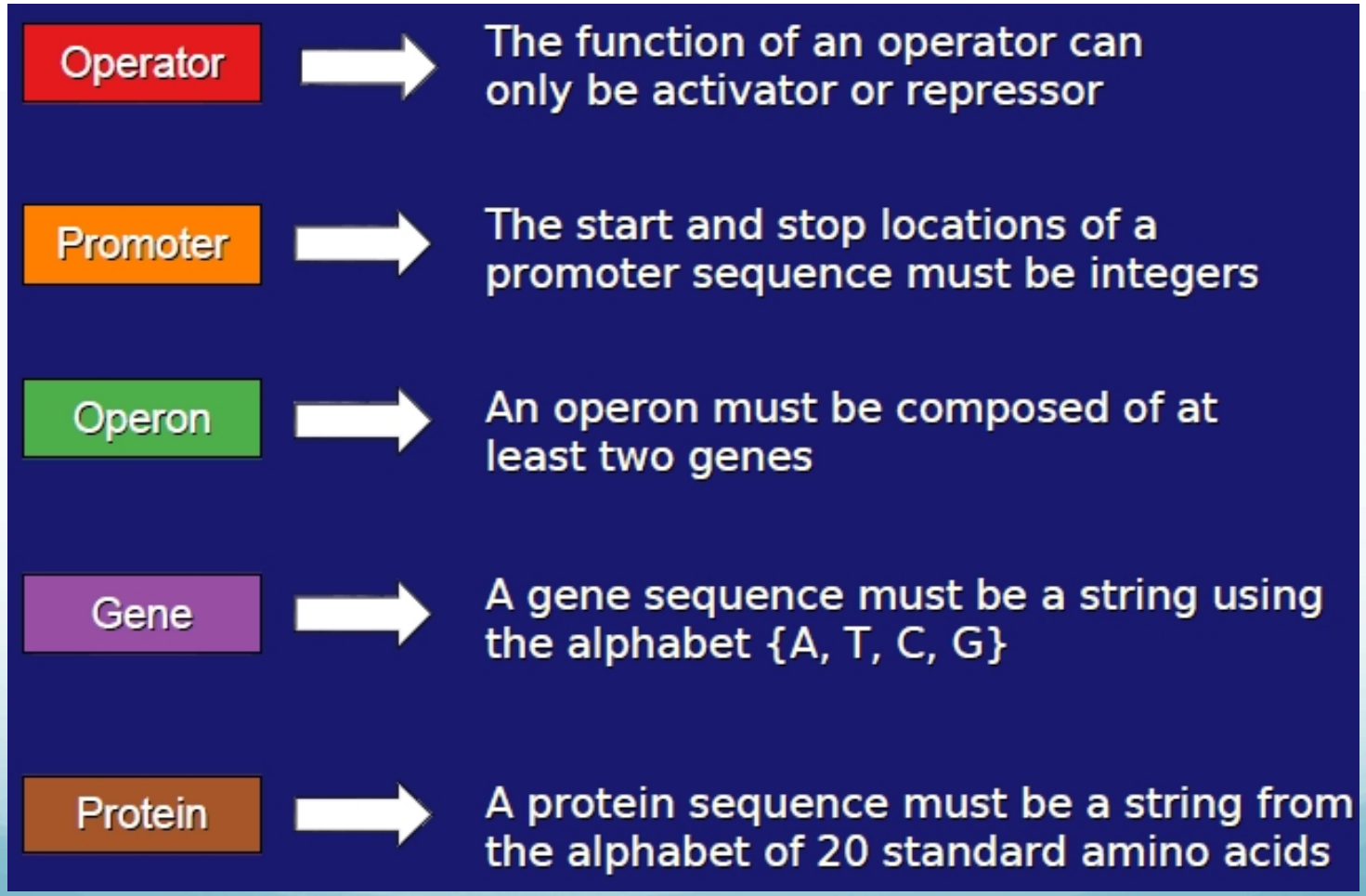
- The terms used to describe the entity are called attributes (adjectives) each type of property is an attribute.



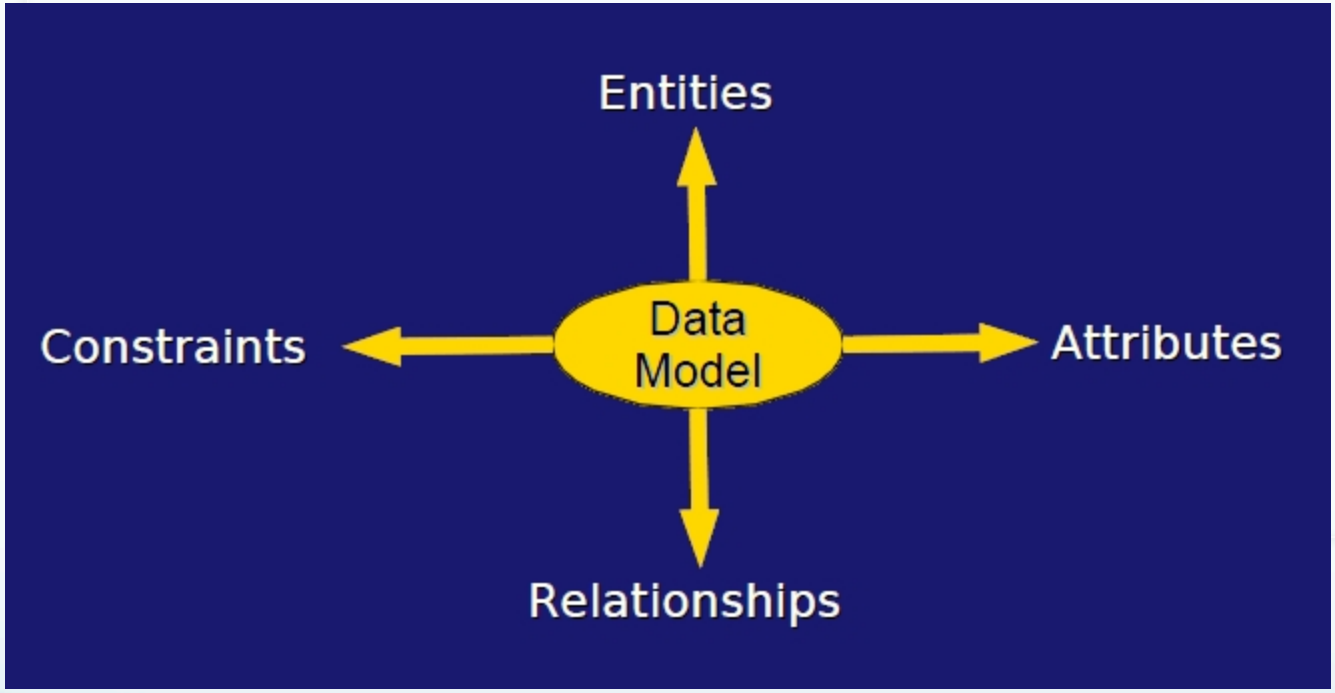
- In complex systems like cells one part often acts upon other parts: the action is a relationship (verb).
- The actions have been divided into 3 types: one-to-one (1:1), one-to-many (1:M) and many-to-many (M:N).



Most attributes are not infinite or continuous – there is a range across which they apply. The range is called a *constraint*.



Getting started : looking at the operon ‘rules’ allows us to propose a data model for storing observations about those properties in specific cases.



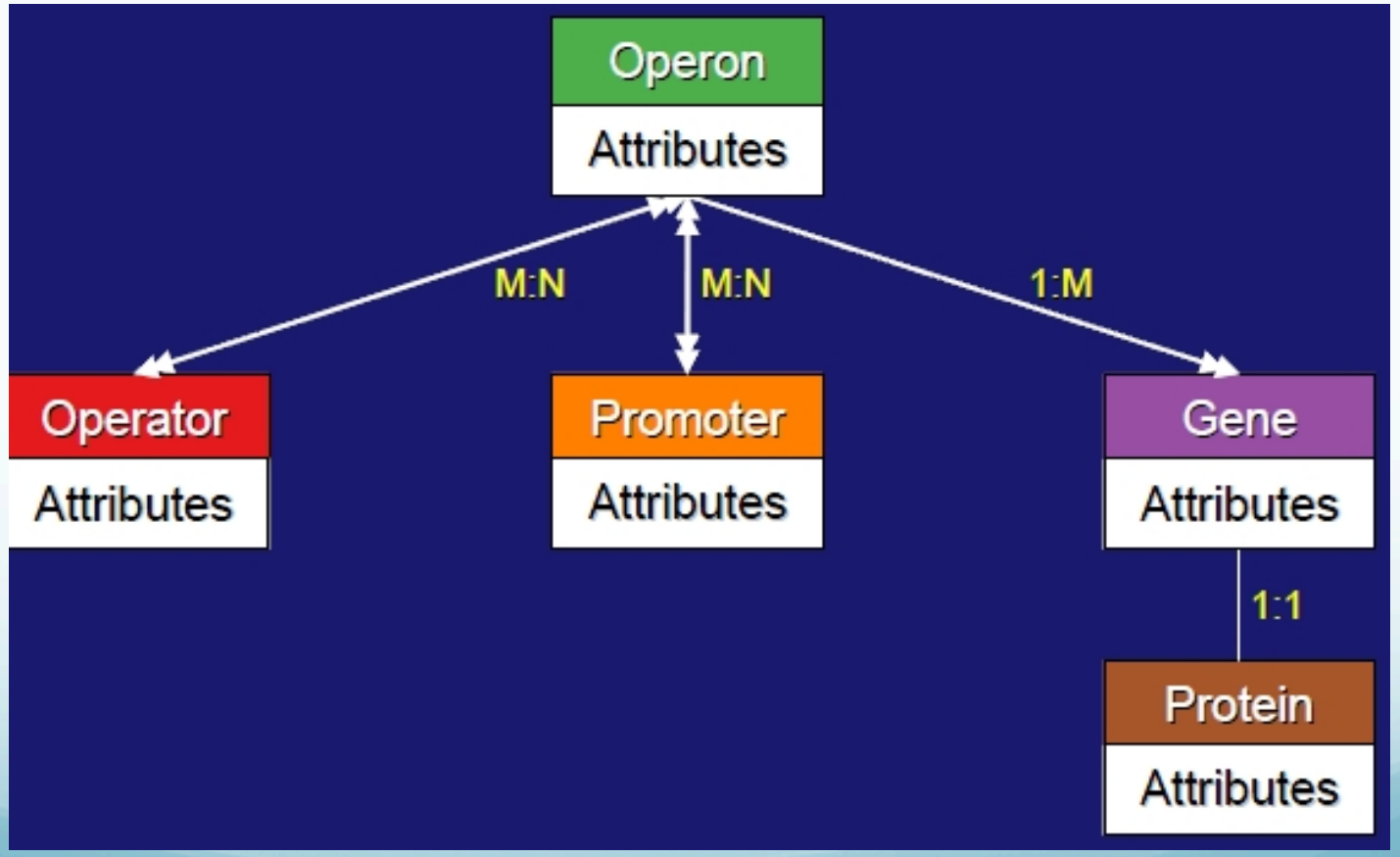
*Business Rules

Business Rule

- Business Rule:
  - brief, precise, and unambiguous
  - Describes policy, principle or procedure
    - Often ascribed to a certain domain
  - *Poorly Named – apply to any organization.*
    - *Think in terms of a lab.*
    - *“PCR machine replicates many sequences.”*

Operon Model

- An ER Diagram – so far.



Operon Model

- A spread sheet is a useful way to lay out the attributes for an entity. Notice that this spread sheet includes 3 of the entities we separated in the diagram. What effect does this have?

- Because we combined entities we have to repeat the values for a lot of the attributes – this redundancy means every times we change one we have to change them all. And some of the attributes are not that relevant to some of the entities – does the start position of a gene matter that much when we are considering properties of a protein?
- If you lose track of one of 5 instances of moaABCDE and change the other 4, now attributes that should be identical will vary.

Operon	Promoter	TSS (Absolute)	TSS (Relative)	Gene	Gene Start Loc	Gene End Loc	Protein GI	Protein Description
moaABCDE	moaAp1	816050	-217	moaA	816267	817256	16128749	molybdopterin biosynthesis protein A
moaABCDE	moaAp1	816050	-217	moaB	817278	817790	16128750	molybdopterin biosynthesis protein B
moaABCDE	moaAp1	816050	-217	moaC	817793	818278	16128751	molybdopterin biosynthesis, protein C
moaABCDE	moaAp1	816050	-217	moaD	818271	818516	16128752	molybdopterin synthase, small subunit
moaABCDE	moaAp1	816050	-217	moaE	818518	818970	16128753	molybdopterin synthase, large subunit
flgAMN	flgAp	1130108	-22	flgA	1129427	1130086	16129035	assembly protein for flagellar basal-body periplasmic P ring
flgAMN	flgAp	1130108	-22	flgM	1129058	1129351	16129034	anti-sigma factor for FliA (sigma 28)
flgAMN	flgAp	1130108	-22	flgN	1128637	1129053	16129033	export chaperone for FlgK and FlgL

We can split up the data to reduce the repeats:

Operon	Promoter	Promoter TSS (Absolute)	Promoter TSS (Relative)
moaABCDE	moaAp1	816050	-217
flgAMN	flgAp	1130108	-22

Sheet 1:
Operon_Promoter

Sheet 2:
Gene_Protein

Operon	Gene	Gene Start Loc	Gene End Loc	Protein GI	Protein Description
moaABCDE	moaA	816267	817256	16128749	molybdopterin biosynthesis protein A
moaABCDE	moaB	817278	817790	16128750	molybdopterin biosynthesis protein B
moaABCDE	moaC	817793	818278	16128751	molybdopterin biosynthesis, protein C
moaABCDE	moaD	818271	818516	16128752	molybdopterin synthase, small subunit
moaABCDE	moaE	818518	818970	16128753	molybdopterin synthase, large subunit
flgAMN	flgA	1129427	1130086	16129035	assembly protein for flagellar basal-body periplasmic P ring
flgAMN	flgM	1129058	1129351	16129034	anti-sigma factor for FliA (sigma 28)
flgAMN	flgN	1128637	1129053	16129033	export chaperone for FlgK and FlgL

The sheets are linked to each other through the Operon column and the values for each instance. Spreadsheet applications may let you document this linkage, otherwise it is up to you to infer its presence.

The information in the spreadsheet is repetitive and redundant – why?

Operon	Promoter	TSS (Absolute)	TSS (Relative)	Gene	Gene Start Loc	Gene End Loc	Protein GI	Protein Description
moaABCDE	moaAp1	816050	-217	moaA	816267	817256	16128749	molybdopterin biosynthesis protein A
moaABCDE	moaAp1	816050	-217	moaB	817278	817790	16128750	molybdopterin biosynthesis protein B
moaABCDE	moaAp1	816050	-217	moaC	817793	818278	16128751	molybdopterin biosynthesis, protein C
moaABCDE	moaAp1	816050	-217	moaD	818271	818516	16128752	molybdopterin synthase, small subunit
moaABCDE	moaAp1	816050	-217	moaE	818518	818970	16128753	molybdopterin synthase, large subunit
flgAMN	flgAp	1130108	-22	flgA	1129427	1130086	16129035	assembly protein for flagellar basal-body periplasmic P ring
flgAMN	flgAp	1130108	-22	flgM	1129058	1129351	16129034	anti-sigma factor for FliA (sigma 28)
flgAMN	flgAp	1130108	-22	flgN	1128637	1129053	16129033	export chaperone for FlgK and FlgL

The 1:M relations require that we enter the same value many times.

Operon Model Expanded 1

- If you want to further simplify any given sheet, the number of sheets will increase – a management challenge ensues.

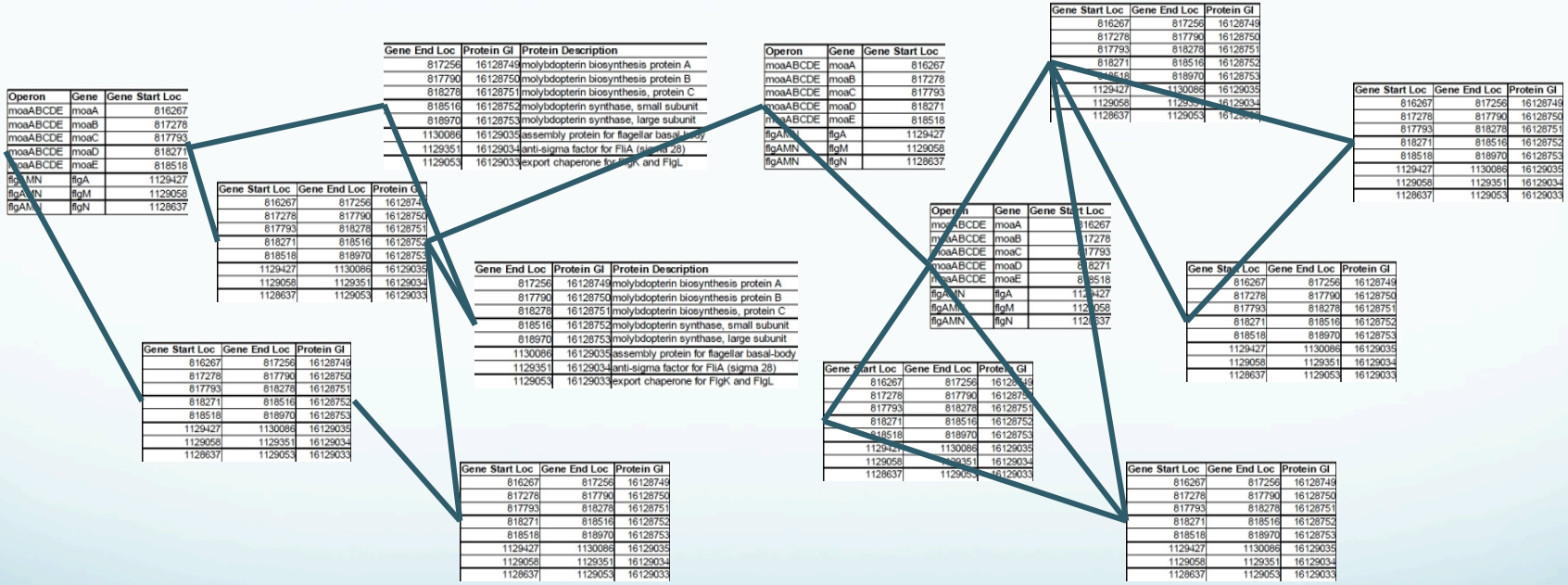
Operon	Gene	Gene Start Loc
moaABCDE	moaA	816267
moaABCDE	moaB	817278
moaABCDE	moaC	817793
moaABCDE	moaD	818271
moaABCDE	moaE	818518
flgAMN	flgA	1129427
flgAMN	flgM	1129058
flgAMN	flgN	1128637

Gene Start Loc	Gene End Loc	Protein GI
816267	817256	16128749
817278	817790	16128750
817793	818278	16128751
818271	818516	16128752
818518	818970	16128753
1129427	1130086	16129035
1129058	1129351	16129034
1128637	1129053	16129033

Gene End Loc	Protein GI	Protein Description
817256	16128749	molybdopterin biosynthesis protein A
817790	16128750	molybdopterin biosynthesis protein B
818278	16128751	molybdopterin biosynthesis, protein C
818516	16128752	molybdopterin synthase, small subunit
818970	16128753	molybdopterin synthase, large subunit
1130086	16129035	assembly protein for flagellar basal-body
1129351	16129034	anti-sigma factor for FliA (sigma 28)
1129053	16129033	export chaperone for FlgK and FlgL

Operon Model Expanded 2

- If you want to further simplify any given sheet, the number of sheets will increase – a management challenge ensues.



As the number of sheets grows you tend to add one more: a Master Plan reminding yourself what the individual related sheets include.

Linking sheets as shown can simplify data entry, but now finding and retrieving data may be very complicated (even with linked tables).

The output of a microarray platform produces 5000 measurements, and the experiment of interest had 21 arrays, triplicates of 7 conditions.

Rows: 21*5000 = 105,000
Storage Feasibility: how many rows does Excel 2007 support?
Retrieval Feasibility:
part 1: locate all raw expression values for one probe across all arrays.
Part 2: How many of the probes produce valid measurements [log10(m) >2.4] across all of the arrays?

What you get: a database is an electronic system for managing data and the meta-data describing that data: properly done, the system is self-describing.

- The management software handles the complex structures, the allocation of files and bits, identification and retrieval tasks.
- The data dictionary holds the meta-data
  - What type of entity is described?
  - What are the attributes of each entity?
  - What are the relationships between the entities?
  - What are the constraints that need to be applied to attribute values and relationships?

Data Records – some vocabulary

Probe_Name	Intensity at 680nm	Intensity at 650 nm	Background at 680nm	Background at 650nm
At_100912	1256	2580	250	126
At_008973	2690	1387	127	95
At_205439	158	128	127	101
At_900196	12876	25908	278	309

- The table format is convenient as an organizing principle
  - Each *field* is a (row x column) intersection with a particular type of information that has meaning, in context
  - A column is a consistent type of information – a type of attribute
  - A row is a record, or an entity instance, every row in this table will show the same number of fields, for different instances, each position in the set of fields will have an attribute of the same type
  - A table is a file containing the collection (entity) of similar records (set of entity instances, or all the rows).

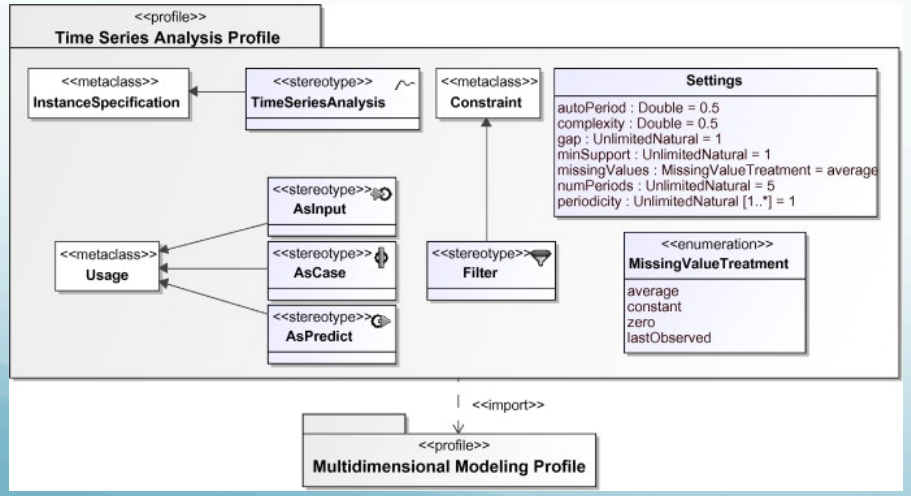
Basic Data Relationships

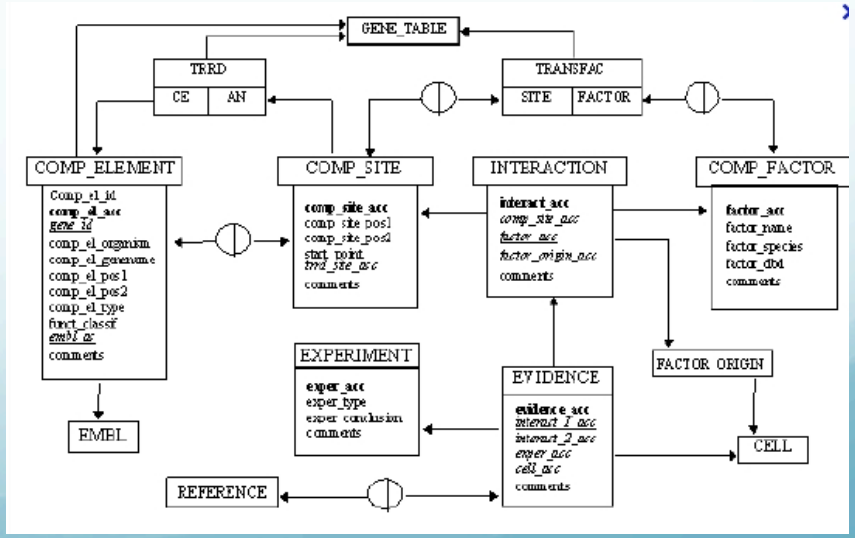
- Determine the *what* – the entities
- List the most important attributes
- Link the entities together with one of the 3 relationship types:
  - One-to-one
  - One-to-many
  - Many-to-Many
- Make sure there are attributes that allow the linking to happen: relationships are stored in the database between entity instances (rows in two tables).
  - Controlled redundancy: the two tables share one type of attribute (one column) in common – this permits logical linkages between tables
    - The columns are declared to contain **keys** in the data dictionary – this tells the application that the tables are linked.

Data Structure Diagram

- A data structure diagram is a representation of a conceptual model that uses standardized graphical notations and terms to document entities and their relationships and transformations.
  - Constraints are also noted

- For database models the entity-relationship (ER) model is common
  - Chen notation or Information Engineering (IE or Crow's foot) notation is most common.
  - UML is also used but can get very complicated





If you repeat information you waste space and possibly cause entry errors, leading to inconsistencies.

Operon	Promoter	TSS (Absolute)	TSS (Relative)	Gene	Gene Start Loc	Gene End Loc	Protein GI	Protein Description
moaABCDE	moaAp1	816050	-217	moaA	816267	817256	16128749	molybdopterin biosynthesis protein A
moaABCDE	moaAp1	816050	-217	moaB	817278	817790	16128750	molybdopterin biosynthesis protein B
moaABCDE	moaAp1	816050	-217	moaC	817793	818278	16128751	molybdopterin biosynthesis, protein C
moaABCDE	moaAp1	816050	-217	moaD	818271	818516	16128752	molybdopterin synthase, small subunit
moaABCDE	moaAp1	816050	-217	moaE	818518	818970	16128753	molybdopterin synthase, large subunit
flgAMN	flgAp	1130108	-22	flgA	1129427	1130086	16129035	assembly protein for flagellar basal-body periplasmic P ring
flgAMN	flgAp	1130108	-22	flgM	1129058	1129351	16129034	anti-sigma factor for FliA (sigma 28)
flgAMN	flgAp	1130108	-22	flgN	1128637	1129053	16129033	export chaperone for FlgK and FlgL

Memory is cheap, but it is a waste to store the same information multiple times.

If you need to change something you must *change every instance* of it, so you have to find every instance of it.

If you lose track of one of 5 instances of moaABCDE and change the other 4, now attributes that should be identical will vary.

It is possible to improve spreadsheets, of course, as shown below.

Operon	Promoter	Promoter TSS (Absolute)	Promoter TSS (Relative)
moaABCDE	moaAp1	816050	-217
flgAMN	flgAp	1130108	-22

Sheet 1: Operon_Promoter

Sheet 2: Gene_Protein

Operon	Gene	Gene Start Loc	Gene End Loc	Protein GI	Protein Description
moaABCDE	moaA	816267	817256	16128749	molybdopterin biosynthesis protein A
moaABCDE	moaB	817278	817790	16128750	molybdopterin biosynthesis protein B
moaABCDE	moaC	817793	818278	16128751	molybdopterin biosynthesis, protein C
moaABCDE	moaD	818271	818516	16128752	molybdopterin synthase, small subunit
moaABCDE	moaE	818518	818970	16128753	molybdopterin synthase, large subunit
flgAMN	flgA	1129427	1130086	16129035	assembly protein for flagellar basal-body periplasmic P ring
flgAMN	flgM	1129058	1129351	16129034	anti-sigma factor for FliA (sigma 28)
flgAMN	flgN	1128637	1129053	16129033	export chaperone for FlgK and FlgL

For this to work you must know (implicit) that the two sheets are related by the 'Operon' column – you find the correct row by using the value in the field in that column.

Data model levels

- Conceptual data modeling, AKA the Use Case – highest level relationships and rules. It includes
  - The scope of the model – WHAT is stored
  - Used for communicating – standard symbols and texts are employed (ER modeling, UML) to developers and users.
  - The processes, such as inputs (and sources), outputs (applications that produce these), transformation steps (for example a reference location based on creating a SAM file)
- Logical data modeling - defines the elements and their characteristics and the relationships that interconnect them.
- Physical modeling - application of the logical data model using database management software (DBMS) – how data are stored.

- Note: the logical data model explicitly determines the structure of data and limits the type of DBMS that will be effective.

If you are sharing the data with other scientists, how do you organize access? How do you prevent others from changing the data?

Security: it is possible to lock fields and sheets to limit accidental changes but this is a very limited approach.

Concurrent access: there is no way for multiple people to use the same sheet at the same time without making separate copies.

Keys

- Key
  - one or more attributes that determines another attribute
  - Used to associate data
  - Rely on determination
    - A→B then A can be used to look up B
- **Primary Key**
  - **Unique reference**
  - **Not Null**
- **Foreign Key -> matches primary key**
- Composite Key
  - Composed of multiple attributes
    - (A, B) → C
- Any attribute that is part of a key is an key attribute by definition