BINF 8211/6211
Design and Implementation of Bioinformatics Databases
Lecture 2

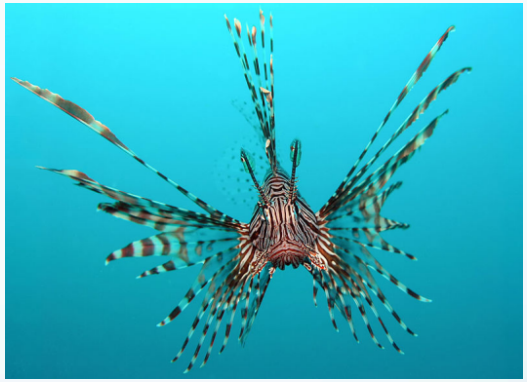
Dr. D. Andrew Carr
Dept. Bioinformatics and Genomics UNCC
Spring 2016

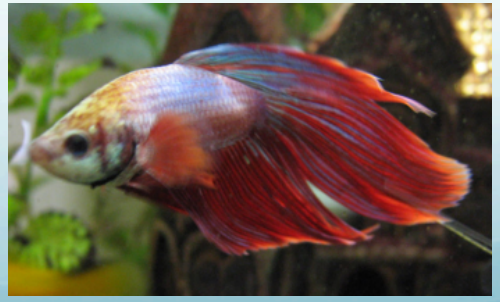
- Scientific Data and Databases
- Data models and Database Management Systems
- Standards: Structured ways to serialize/ exchange data, shared core models, Ontologies

There are 3 domains to master

Database design considerations

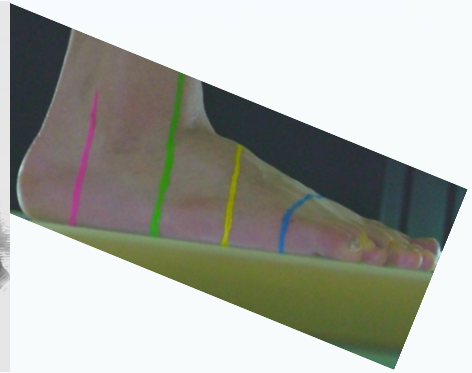
- What level of detail is stored?
  - How much metadata?
  - How complex is the data?
  - Do all the entries have the same number of attributes?
    - What do you do about missing information?
    - How do you account for the variation?
      - Individual tables for different types of records or one large table with lots of attributes?
  - Are all the measures on the same scale?
  - What is the resolution?

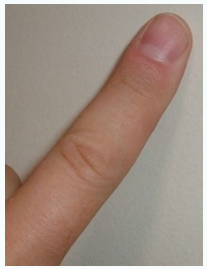




Scientific data comes from measurements and controlled observations.

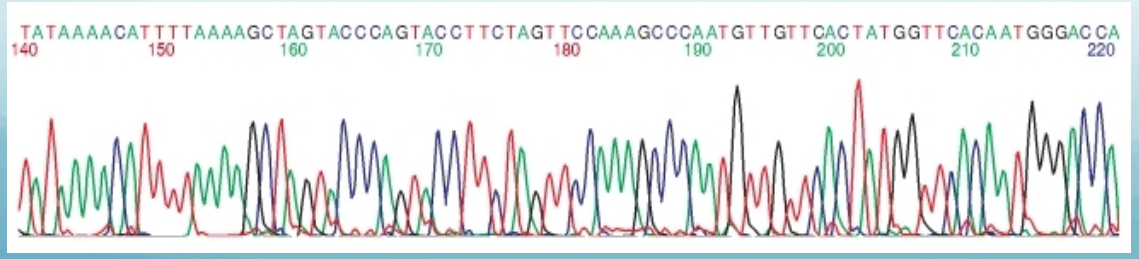
- Instruments have limitations
  - Precision, accuracy, uncertainty, **units**

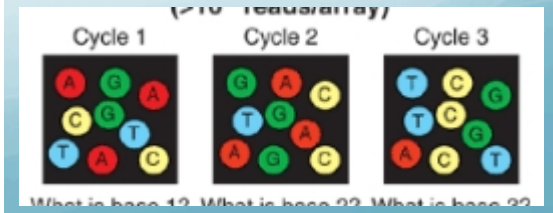




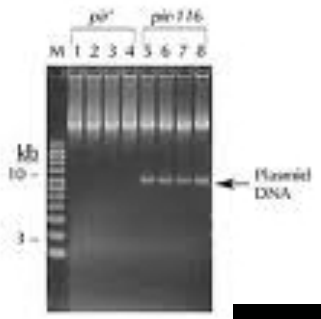
* 10 or 12? * 16?

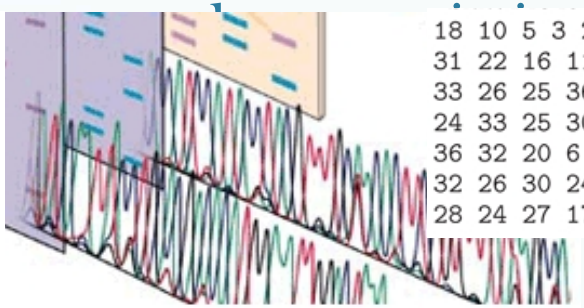
- And…technology changes



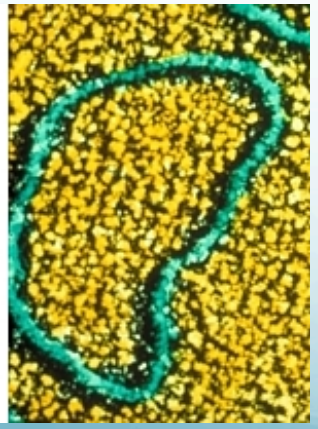


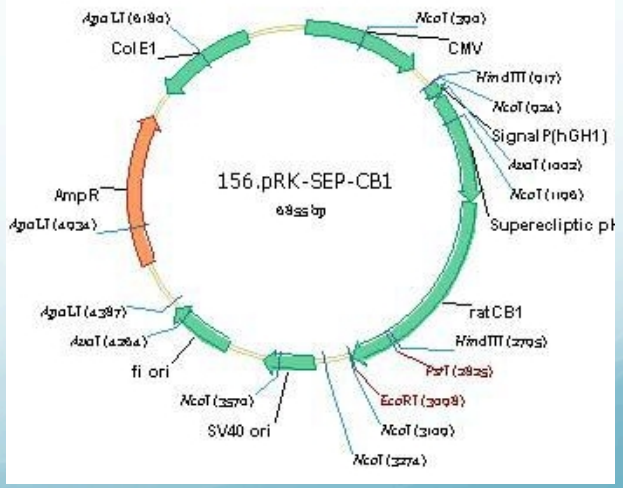
Measurements have *types* (distinct from data types).





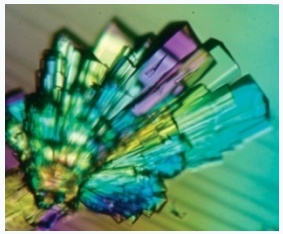
18 10 5 3 2 1 1 1 1 1 1 1 1 1 1 22 37
31 22 16 11 6 1 26 34 30 11 33 26 30 21
33 26 25 36 32 16 36 32 16 36 32 20 6
24 33 25 30 25 2 24 36 32 15 35 31 17
36 32 20 6 25 29 20 30 25 4 32 26 32 23
32 26 30 24 33 26 35 31 14 28 27 30 22
28 24 27 17 32 23 28 28

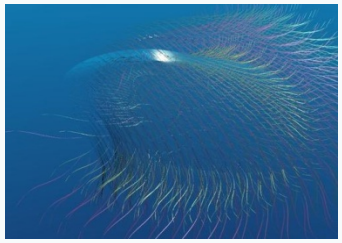




The context of measurements includes assay type, and value attributes.

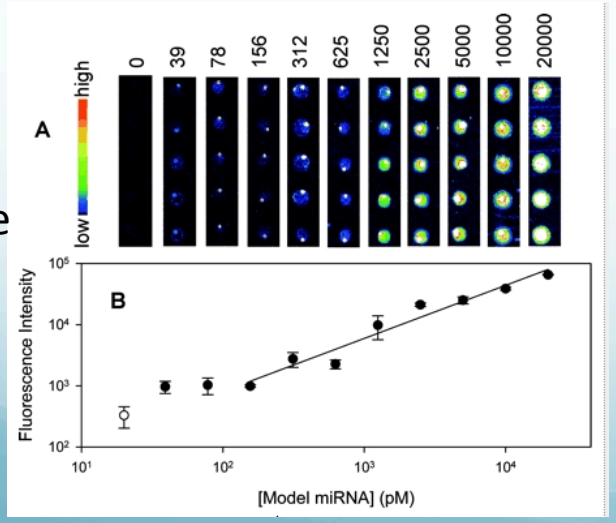
- Values have scale and range (and units and sources)
  - The range has boundaries
  - The scale need not be linear
  - What information is lost or confused when multiple dimensions are flattened?





Scale?

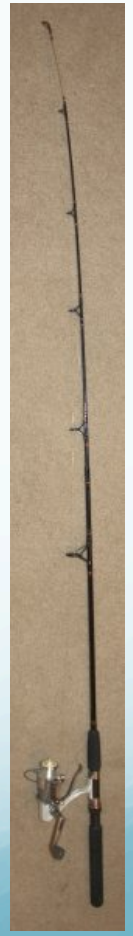
Log scale



A model simplifies

Database design considerations

- How is the data going to be accessed?
  - What is the most commonly sought information?
    - Is it a single item? Or is it many items?
    - How big is it?
  - Common of method of access
    - Web Browser
    - Specific database application
      - Is the data on a single workstation or is it to be accessed remotely?
  - Data security
    - HIPPA compliancy?
  - Which is the right tool for the job?



So what is this data of which you speak?

Thorough understanding of the sample, the technology and questions that the original researchers were asking.

What meta-data would be applied?

What information could be associated with the data?

Discovering Knowledge: Integration

- Data
  - What is being measured, how is it labeled?
  - Is the file format self-describing?
  - If there is a source database, what filters are used?
- Annotation vocabulary
  - What type of structure is used? (ontologies)
  - Has the ontology been logically tested? (e.g. Protégé)
  - What application has been used to define and work with the ontology (OWL/RF/RDFS)
- Representation and Manipulation
  - How are data elements described and connected? (models)
  - What type of database application has been chosen and how does that affect data structures? (DBMS)
  - What data definition language and data manipulation language are being used? (SQL)
  - How unique is each data element (do I have to look in more than one place and will those places be synchronized)? (Normalization)

Database History- flat files

- Original system were file based.
  - File System based.
  - Record management
  - 1960s

- A flat file contains
  - records but no structured relationships
  - a fixed number of fields that may not be formatted.
- File format:
  - Is this an text file or a binary file
  - How is the information organized
  - For microarrays, examples include .DAT, .CEL, .CHP, .TXT. .RPT, .EXP

- Data types
  - Integers
  - Booleans
  - Characters
  - Floating-point numbers
  - Alphanumeric strings
  - Categorical data (enumerated)
  - Array data

Marker	CB26-2	CB511	CB512	CB1171	CB1193	N 4-13	N 4-14	N 4-7	N1 4-21
33540_at: (---) ---	4.98E1	2.30E1	1.20E1	2.80E1	5.91E1	5.17E1	5.75E1	3.43E1	8.10E1
33541_s_at: (LAIR2) leukocyte...	9.05E1	7.35E1	7.15E1	8.47E1	3.03E1	1.34E2	9.56E1	4.87E1	1.72E1
33542_at: (ZNF280B) zinc fing...	8.00E0	1.10E1	4.31E1	5.70E0	4.03E1	7.30E0	1.65E1	4.75E1	4.61E1
33543_s_at: (PNN) pinin, des...	7.26E2	6.99E2	6.87E2	5.65E2	9.12E2	5.42E2	5.20E2	5.22E2	4.07E2
33544_at: (UNC5C) unc-5 ho...	2.17E1	2.43E1	5.30E0	8.60E0	3.70E0	4.30E0	1.35E1	4.10E0	3.60E0
33545_at: (SCN4A) sodium ch...	4.39E2	5.78E2	4.42E2	3.70E2	2.90E2	2.99E2	2.86E2	2.44E2	2.92E2
33546_at: (SPRR1A) small pro...	2.40E2	3.36E2	2.90E2	3.00E2	1.91E2	1.35E2	9.60E1	6.98E1	7.27E1
33547_i_at: (FUT2) fucosyltran...	6.50E0	8.00E-1	1.40E0	7.00E-1	1.50E0	1.35E1	1.40E0	1.60E0	2.20E0
33548_f_at: (FUT2) fucosyltran...	2.90E0	9.00E-1	2.20E0	4.10E0	5.00E0	1.50E0	3.40E0	9.00E-1	1.30E0
33549_at: (BDKRB1) bradykini...	3.19E1	1.54E1	4.70E0	1.26E1	9.00E-1	4.50E0	1.23E1	2.86E1	2.81E1
33550_at: (HTR2C) 5-hydroxytr...	2.47E1	2.08E1	1.52E1	2.84E1	1.43E1	3.52E1	1.69E1	6.80E0	6.80E0
33551_s_at: (HTR2C) 5-hydrox...	1.70E0	2.20E0	2.30E0	1.10E0	9.00E-1	8.60E0	4.20E0	5.40E0	1.38E1
33552_at: (SLCO1A2) solute c...	6.90E0	3.10E0	4.80E0	3.80E0	3.80E0	3.20E0	1.19E1	4.30E0	3.19E1
33553_r_at: (CCR6) chemokin...	6.37E1	7.74E1	1.07E2	8.03E1	6.28E1	4.70E2	4.87E2	4.17E2	4.00E2
33554_at: (LOC100129973) hy...	1.03E1	3.02E1	2.21E1	4.27E1	2.87E1	4.77E1	2.00E1	1.60E1	2.41E1
33555_at: (LILRA4) leukocyte i...	5.69E2	5.93E2	6.95E2	3.57E2	5.57E2	2.17E2	2.22E2	2.53E2	2.81E2
33556_at: (FICD) FIC domain c...	1.16E1	2.65E1	6.90E0	1.88E1	5.60E0	4.61E1	3.97E1	1.66E1	9.00E0
33557_at: (C22orf31) chromos...	1.38E1	1.61E1	4.60E0	8.30E0	4.30E0	4.40E1	4.77E1	7.60E0	1.46E1

Flat Files

SAM Alignment Tags

```
NM:i:0  SM:i:37 AM:i:0  X0:i:1  X1:i:0  XM:i:0  XO:i:0  XG:i:0  MD:Z:37
XT:A:R  NM:i:0  SM:i:0  AM:i:0  X0:i:4  X1:i:0  XM:i:0  XO:i:0  XG:i:0  MD:Z:37
NM:i:2  SM:i:0  AM:i:0  X0:i:4  X1:i:0  XM:i:0  XO:i:1  XG:i:2  MD:Z:18^CA19
SM:i:0  AM:i:0  X0:i:5  X1:i:0  XM:i:0  XO:i:1  XG:i:2  MD:Z:35
```

Field	Alignment 1	Alignment 2
QNAME	1:497:R:-272+13M17D24M	19:20389:F:275+18M2D19M
FLAG	113	99
RNAME	1	1
POS	497	17644
MAPQ	37	0
CIGAR	37M	37M
MRNM/RNEXT	15	=
MPOS/PNEXT	100338662	17919
ISIZE/TLEN	0	314
SEQ	CGGGTCTGACCTGAGGAGAACTGTGCTCCGCCTTCAG	TATGACTGCTAATAATACCTACACATGTTAGAACCAT
QUAL	0;==-==9;>>>>>=>>>>>>>>>>>>>=>>>>>>>>>>	>>>>>>>>>>>>>>>>>>>><<>>><<>>4::>>:<9
TAGs	XT:A:U NM:i:0 SM:i:37 AM:i:0 X0:i:1 X1:i:0 XM:i:0 XO:i:0 XG:i:0 MD:Z:37	RG:Z:UM0098:1 XT:A:R NM:i:0 SM:i:0 AM:i:0 X0:i:4 X1:i:0 XM:i:0 XO:i:0 XG:i:0 MD:Z:37

MD	Z	String for mismatching positions. *Regex*: [0-9]+(([A-Z]|\^[A-Z]+)[0-9]+)*[6]
MQ	i	Mapping quality of the mate/next segment
NH	i	Number of reported alignments that contains the query in the current record
NM	i	Edit distance to the reference, including ambiguous bases but excluding clipping

History of Data Models– The Model

- Flat files are not sufficient – Need a better way.
- STOP –
  - How is a model constructed???
  - What are the basic building blocks.

Data Model Building Blocks

- **Entity**:
  - The most basic level of the model.
  - Person, place, thing, event (fish)
    - Row in a table.
- **Attribute**
  - “Characteristics of an entity.” Rob and Coronel
    - Name, color, weight, height, fresh water….
    - Column in a table.
- **Relationship**
  - Association between entities
  - Tank and Fish: The fish lives in the tank.

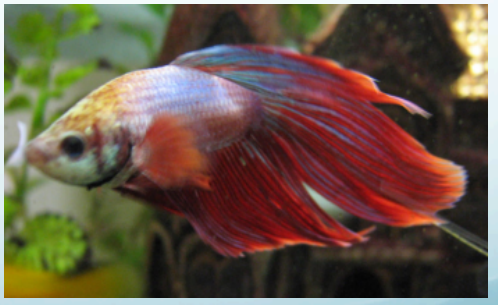
Data Model Building Blocks Relationships

- One to Many
- Many to Many
- One to One
- One to (Finite)
  - Constrained relationship

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."*

- " Each beta has its own tank."



History of Data Models

TABLE 2.1 Evolution of Major Data Models

GENERATION	TIME	MODEL	EXAMPLES	COMMENTS
First	1960s–1970s	File system	VMS/VSAM	Used mainly on IBM mainframe systems Managed records, not relationships
Second	1970s	Hierarchical and network	IMS ADABAS IDS-II	Early database systems Navigational access
Third	Mid-1970s to present	Relational	DB2 Oracle MS SQL-Server	Conceptual simplicity Entity relationship (ER) modeling support for relational data modeling
Fourth	Mid-1980s to present	Object-oriented Extended Relational	Versant VFS/FastObjects Objectivity/DB DB/2 UDB Oracle 10g	Support complex data Extended relational products support objects and data warehousing Web databases become common
Next Generation	Present to future	XML	dbXML Tamino DB2 UDB Oracle 10g MS SQL Server	Organization and management of unstructured data Relational and object models add support for XML documents

History of Data Model - Standards

- Data serialization and exchange – retaining context while changing models
- Ontologies: standardized vocabularies with logical relationships as part of the definition
- Generic Model Organism Databases + Tools

Tools for the Course

- SQL Browser
  - Platform independent
  - SQL Lite based
    - May not work for everything due to size
    - http://sqlitebrowser.org/
- SQL Lite
  - https://www.sqlite.org/index.html

In class assignment #1

- Create and Populate 2 Tables with data from the Web.
  - EMBL or NCBI
    - Protein or Gene sequence
    - One Table contains Sequence Information
    - Each table needs to have the following considerations
      - How many attributes?
      - # of rows 2-3
    - Due to Dr. Carr via email by 8 a.m. on Thursday January 21st 2016.
    - email me the .db file with proper labeling.