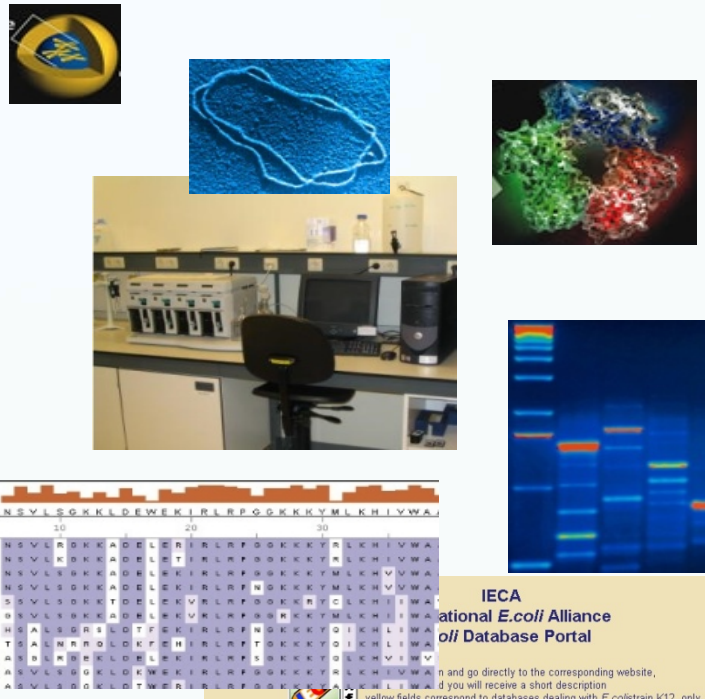
BINF 8211/6211
Design and Implementation of Bioinformatics Databases

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

What can you expect from this course, this semester?

Where	What
BINF 105	8211 6211

Who	When	Policy, Grades, etc →
Carr	TTh 8:00 am	



- Class Conduct
- Course Overview (theory and practice)
  - Data (production, file formats, sources)
  - Ontologies + OWL + Protégé
  - Data Models + Databases + SQL + Normalization
- → INTEGRATION = New Information→ Knowledge

*Modes of Communication

- Professor: Dr. D. Andrew Carr

  Office: Bioinformatics 353??

  email: dcarr10@uncc.edu
  Office Hours: By Appointment.

  TA: Maoxuan Lin – hours will be posted

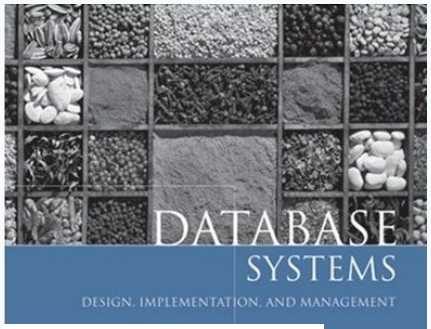
Why is there no textbook?

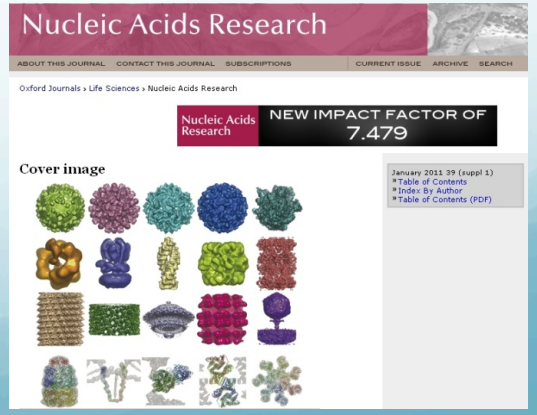
Recommended resources/references:

- Textbooks:
  - Rob and Coronel, ed 8, “Database Systems: Design, Implementation and Management” (2009)
  - Possibly: Robinson and Bauer “Introduction to Bio-Ontologies” (2011)

- Journals:
  - Nucleic Acids Research – January edition: http://nar.oxfordjournals.org/content/39/suppl_1.toc
  - Oxford Press: Database - http://database.oxfordjournals.org/content/current
  - Papers will be posted on Moodle

- eTutorials: various, examples will be posted.





Your time investment is essential if this class is to be useful.

- Methods involve active skills, requiring practice.

  Concepts + Tools = {null}

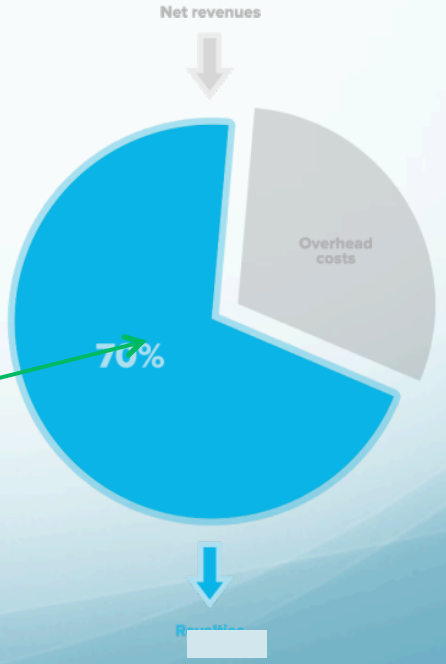
  Concepts + Tools + Practice = {Competence}

- Come prepared, don't wait until the last weekend to work on assignments.

- In addition to active skills, critical thinking, communication, ethical practice are important

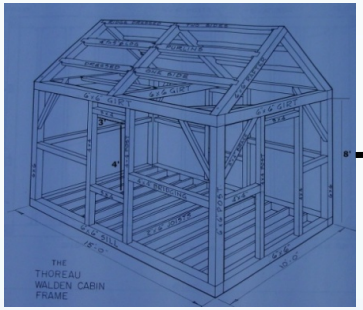
- Your work must be your own, but citing authority is essential – do so properly!

  How much must be 'your own'?



Assessment will be on how well you master the theory, the technical vocabulary and how well you put the concepts into practice.

- Problem sets: are of limited scope to practice basic skills
- Homework: integrates concepts with the end goal of helping you develop the elements needed to implement your database.

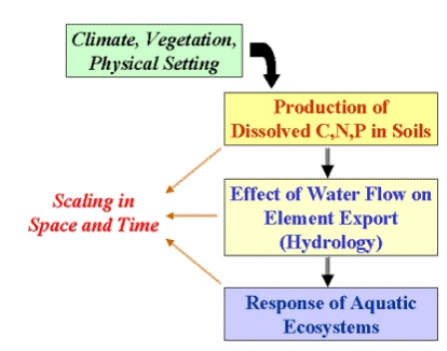


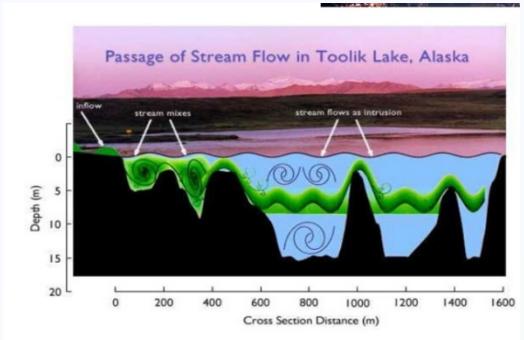
- Exams: midterm and final
  - Comprehensive on concepts covered in assigned reading material and lectures
- Project

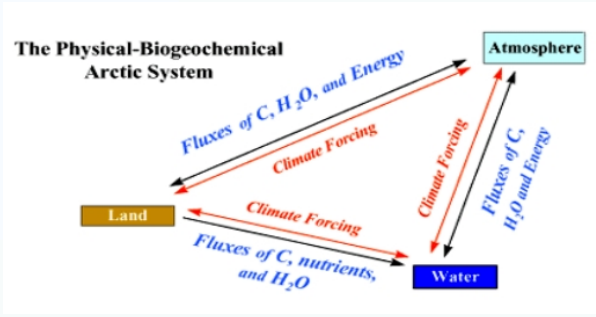
Grading proportions on class work

- Homework: building blocks of your Project – cumulative
  - 15%
- Problem Sets: indicate progress/mastery of basic skills
  - 15%
- Quizzes:
  - 10% -- may be given at random
- Exams: these indicate your knowledge of theory, sources, content
  - 15% Midterm
  - 15% final
- Project: indicates your ability to integrate theory and practice with skill.
  - *Presentation: show your work live and convince us of its importance - 15%*
  - *Paper: present the structure and give a deeper understanding of what went into the design and implementation – 15%*

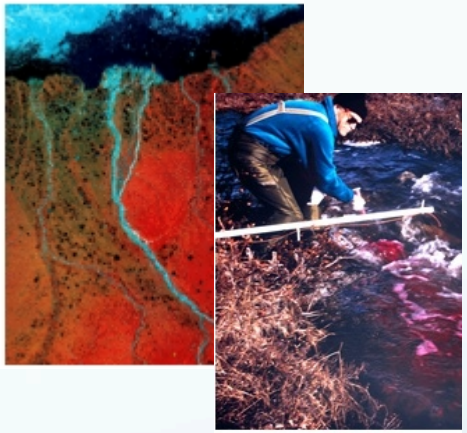
Homework assignments: the building blocks for your Projects.

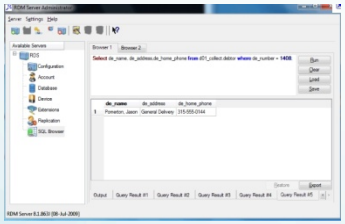


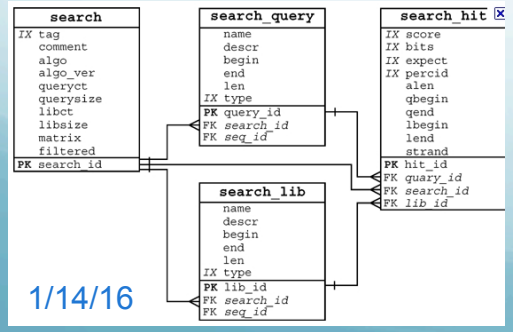




Questions + Data+ Design, iterate.
The feedback you get at each stage will be important to your final success.

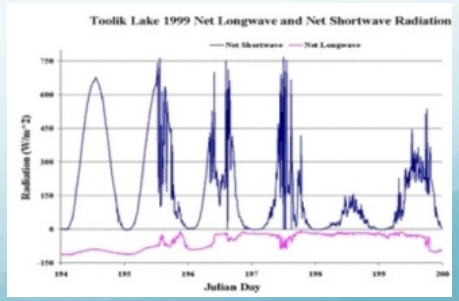






Location	Ecosystem Type	Measurements	Observations (in red) Experiments (in blue) Synthesis (in green)
Terrestrial Experimental plots	Moist Acidic Tussock, Wet Sedge, Nonacidic Tussock	Soil water chemistry, C and nutrient production Water additions to tundra $^{14}CO_2$ labeling	
Tussock Watershed	Moist Acidic Tussock, Primary Stream	Stream flow and chemistry, rain events Soil water chemistry Hydrology and biogeochemistry model	
Inlet Series of Lakes in the Toolik Basin	Lakes and Streams	Lake and stream chemistry Lake mixing and primary production Integration of ecosystems across the landscape Hydrology and biogeochemistry model	
Toolik Lake, Lake E5	Lakes and their inlet streams	Ecological and chemical impacts of storm events (major inflows) on lakes	

Land-Water Site Codes LTER site codes for Land-Water sites

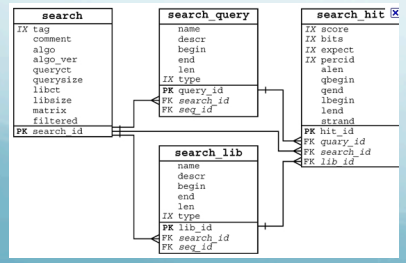


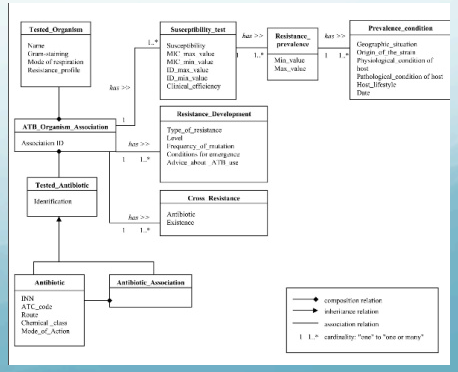
Your project is intended to be: challenging, reflect current scientific research, and to result in a real product for your portfolio.

- General field : (optional Bioinformatics system)
  - Manage raw data and derived data
  - Map to a Model Organism Database basic schema
  - Use at least 2 bio-ontologies
- Teamwork: you *may* work in teams of 2, you do not *have* to.
  - You must have one part of the project that is yours, to be presented independently.
  - Each team member must give a separate presentation, with unique slides
  - Each team member must produce an individual project report
  - Each team member is responsible for the success of his project – if your partner is not carrying her weight the professor is not going to mediate for you or forgive you some of the requirements.

Formal Project requirements

- Develop 5 non-trivial scientific questions for which your database immediately returns results.
  - Which sequence variants result in amino acid coding changes?
- Produce an annotated model in an accepted format for each implementation stage.
  - Conceptual model (also called process or workflow diagram)
  - Logical model (specialized to the entity-relational framework)
  - Schema/normalization (prepared for implementation and tested for redundancy)
  - Implementation and data population
- Annotation: fully describe the meaning of the symbols and content
- Comply with standards: explain data source(s), provenance, validation tests, how you determined that values could be compared from the different measurement platforms.
- Provide logical and SQL queries that retrieve data to answer the questions you have posed. Show how you have tested their accuracy.





Project: how much work is enough? or what is way too ambitious?

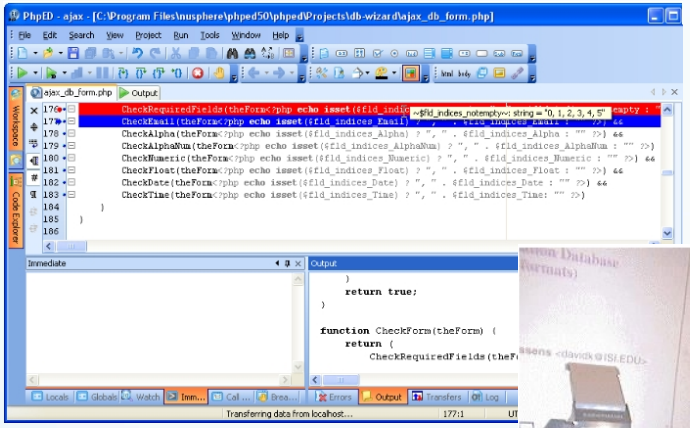
- Model Complexity
  - The schema must include
    - At least 10 data tables and no more than ~20 non-linking tables (note linking tables add tables and functionality but are not design-heavy)

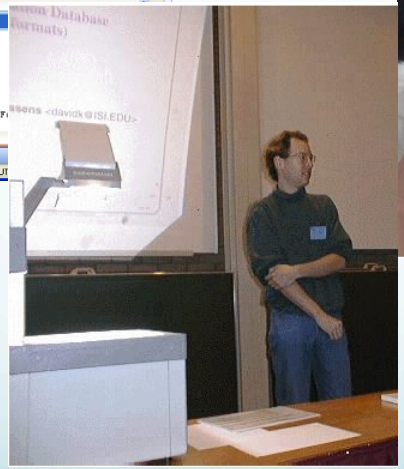
- Data volume
  - The data size must be such that 2 of your tables have >10,000 rows of data
    - Don't aim for more than 1,000,000 rows unless you have programming abilities
    - All data sources must be explicitly referenced
    - Note that access to some data repositories may require registration or writing for permission to a database (so there may be a time delay).

- Query complexity
  - The **queries** must include 4 that require at least 2- table joins and each join must differ from the others.
  - There must be 1 query that requires a 3-table join.

- Testing: you must show that you have a plan and used it
  - Testing data integrity (upload)
  - Query accuracy

- Presentation: you have 15 minutes to introduce your scientific domain and demo your project.
  - Keeping to time allotted counts!
  - Your peers will participate in assessment (how effective are the slides, how well did your queries run, etc.)





15 minutes

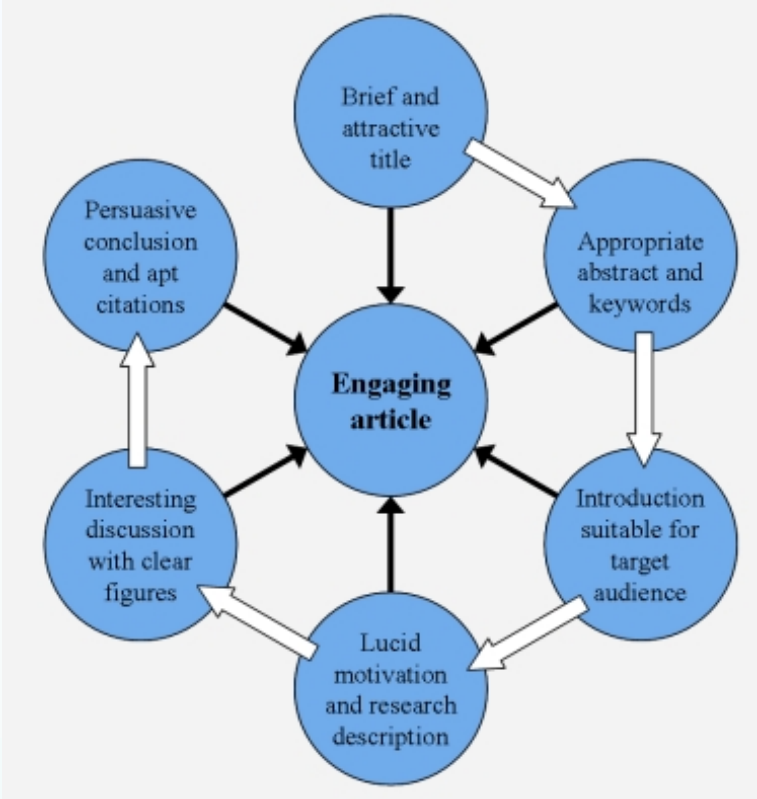
Specific Requirements for Project Presentation

- Project Presentation
  - Length: 15 minutes TOTAL - including questions, set-up time etc.
  - Demo Preparation: one functional database, preloaded on a laptop or server, be ready to cut and paste the queries from a text file.
  - Context Preparation: 10-12 slides (submitted the day of the presentation) that cover the following:
    - Background: scientific context, current data repositories with their strengths and weaknesses.
    - Introduction: what are the 5 questions, why are they important.
    - Methods: logical model and decisions made when transforming into the schema, data sources, upload and validation method
    - Results: run the queries live, show the results – you can also provide summaries or graphs if they inform the outcome.
    - Discussion: why are the results significant?
  - The class will assess each presentation, delivery and style will count along with content, and ability to answer 2-3 audience questions.

Project Report

The project *report* lets you show how much work you did and how clever you have been, in more detail than the presentation allows.

The **style** and **format** must adhere to professional technical writing standards– **use journal reading assignments as your guide.**



Good writing counts.

See http://www.enago.com/blog/how-to-engage-the-attention-of-the-reader/ for suggestions.

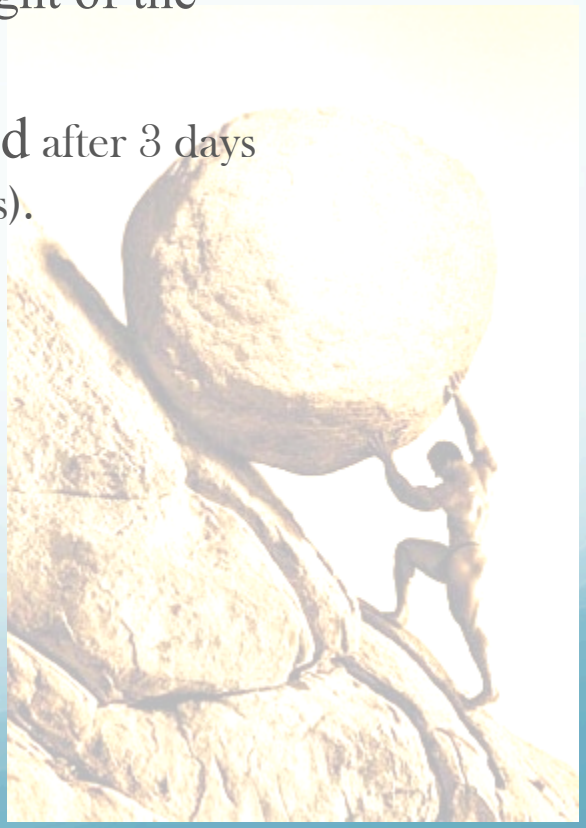
Specific Requirements for Project Reports

- Project Reports – hard copy.
- Due May 5th, 5pm, to Dr. Carr.
  - 8-10 pages (single spaced), written and formatted for a journal like the Oxford 'Database' (see http://database.oxfordjournals.org/) journal.
  - You must have:
    - Title/Author page
    - Abstract
    - Introduction and Background
    - Materials and Methods
    - Results
    - Discussion and Conclusions
    - References (at least 10)
    - Supplementary Materials (on a CD: data, scripts, SQL, etc.)
- You should give me enough material that I can replicate your database and your results. This includes and loading scripts.

Where to Find Assignments

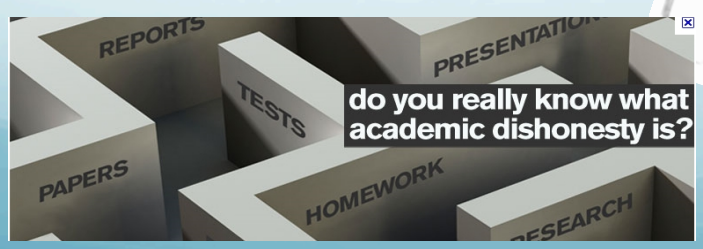
- Homework and Problem Sets will be announced in class
  - Online: Moodle? Website?
  - Due dates will be posted when they are assigned

- Post due assignments to Moodle, no later than midnight of the posted due date.
  - Late = 10% lower score **per day**, not accepted after 3 days (barring the usual exceptions for illness and emergencies).
  - *Your name must be part of the file name*!
  - BINF6211s2016_HW1_CARR



Academic Integrity

- All UNC Charlotte students have the responsibility to be familiar with and to observe the requirements of the UNC Charlotte Code of Student Academic Integrity. This Code forbids
  - Cheating
  - Fabrication or falsification of information
  - Multiple submission of the same work for different assignments
  - Plagiarism, abuse of academic materials
  - Complicity in academic dishonesty (helping others to violate the Code).
- Violators risk being expelled permanently from UNC Charlotte and having this fact recorded on their official transcripts.
- Penalties
  - Zero credit on the work and substantial reduction of the course grade, generally to an 'F'.
  - Relevant documents include the online Handbook at http://www.legal.uncc.edu/policies/ps-105.html, "Settlement of a Charge of Academic Dishonesty Form" and the essay "UNC Charlotte Student Academic Integrity: On Deciding Guilt and Punishment," by R.H. Toenjes.



Data – Databases Design Reality?!?

Dr. D. Andrew Carr

Database 821

Jan 12, 2016

Guest Lecture

What is Data?

- "Data are raw facts" ~Rob and Cornel pg. 5
- What are examples of raw data in Bioinformatics?
  - How complicated can a single data point be?
  - How important is the context of the data point?
  - Examples:
    - In a crystal structure: X,Y,Z atom type, SOF
      - What else would you want to know?
    - X, Y, signal intensity, STDV, spot area, probe ID, probe sequence
      - What is missing, what else would you want to know?
- Information is knowledge added data
  - "Processed data"

Protein: 1GBG

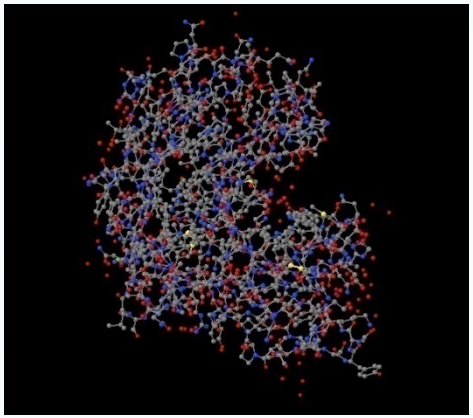


Image above produce using JMOL

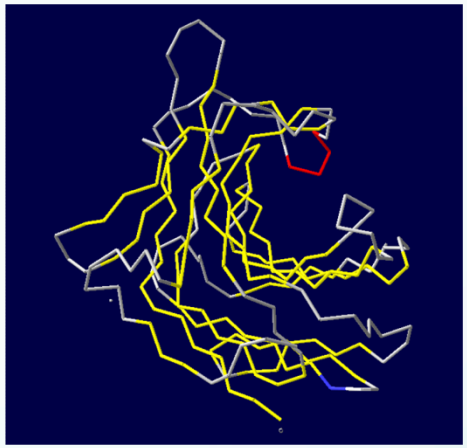


Image above produce using GLISTEN

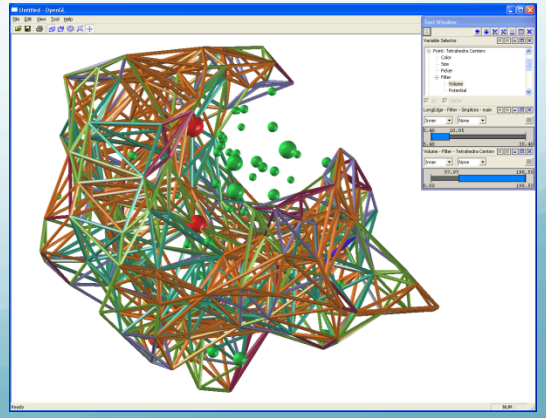
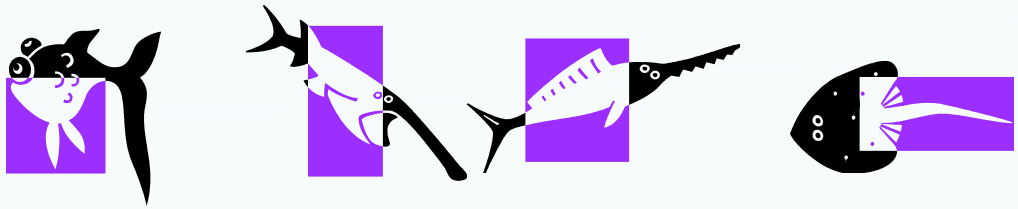


Image above produce using GLISTEN

What is a Database?

- Given a set of items….



- Is this a database?

Images extracted from Windows Clipart.

What is a Database?

- In the most generic terms a database is an organized collection of information.

- Why does it need to be organized?

What is database design?

- Database design at its most basic level is the logical manner in which the data/information is organized and stored.

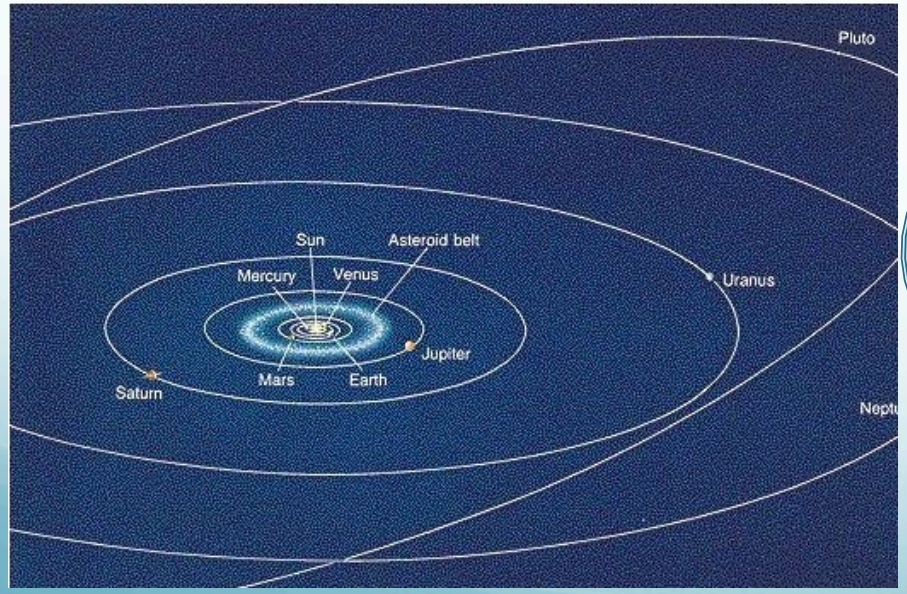
- So, why is database design important?
  - What are the basic design considerations be?

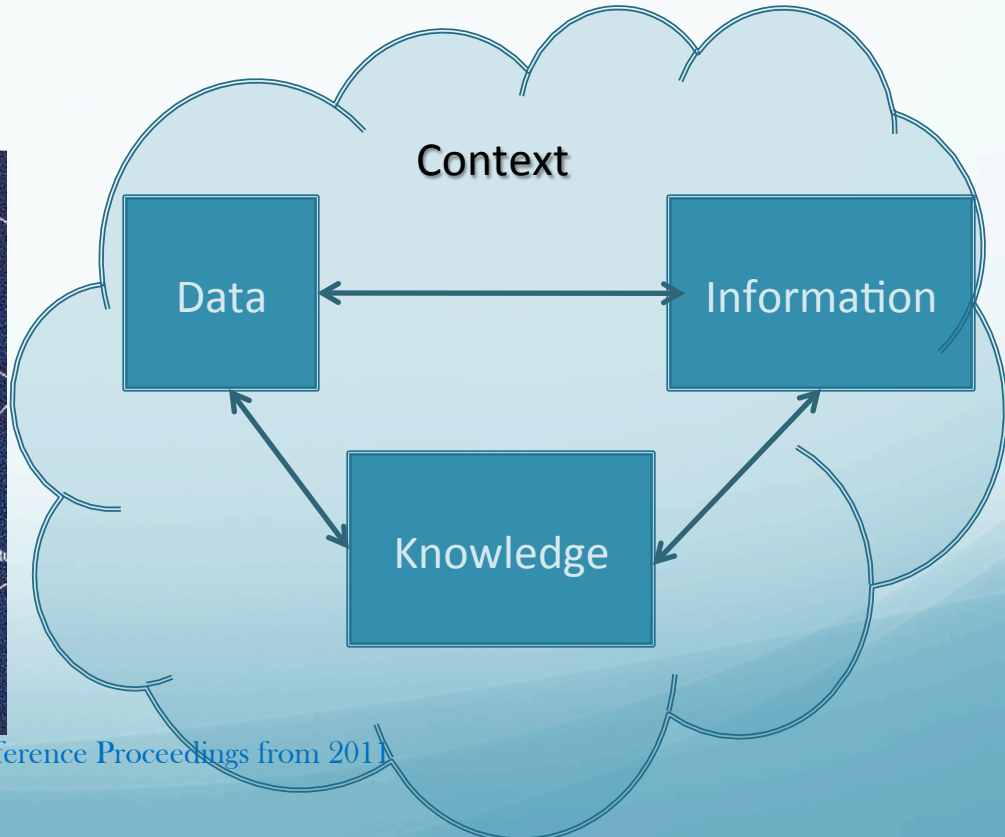
AFFX-BioC-5_at	153.71754	116.37271	72.11234	89.12201	54.45521	95.60325	85.12175
AFFX-BioC-3_at	92.98369	64.34562	45.73359	56.8517	50.3705	66.33909	64.24095
AFFX-BioDn-5_at	53.81335	54.59389	54.42901	62.13527	43.32661	65.01606	63.50323
AFFX-BioDn-3_at	43.32732	56.9259	81.80589	80.48455	55.18736	82.[illegible]	85.[illegible]
AFFX-CreX-5_at	487.10306	0.00268	3.33077	0.00239	657.22284	831.09113	0.00129
AFFX-CreX-3_at	151.4509	1.1859	1.92758	1.82695	303.4081	488.[illegible]	0.01253
AFFX-DapX-5_at	15852.8916	9163.0166	11118.37598	10364.0869	11229.052[illegible]	9227.37598	[illegible]
AFFX-DapX-M_at	15646.11719	8815.02734	11492.65723	11115.4902	11173.5	[illegible]	[illegible]
AFFX-DapX-3_at	13327.5166	7178.13818	8905.83496	7994.83057	8954.11[illegible]	[illegible]	[illegible]
AFFX-LysX-5_at	4416.71094	1821.91516	2191.10815	1771.99817	2122.41[illegible]	[illegible]	[illegible]
AFFX-LysX-M_at	3948.22827	1361.40381	1813.38831	1192.94067	1539.9738[illegible]	[illegible]	[illegible]
AFFX-LysX-3_at	1869.84033	805.40814	917.98145	982.30621	1243.49365	[illegible]	[illegible]
AFFX-PheX-5_at	8236.10938	3518.94482	4350.33154	4205.71045	4992.44678	[illegible]	[illegible]
AFFX-PheX-M_at	7138.53467	2829.36182	3929.41772	3350.14233	4006.66211	27[illegible]	[illegible]
AFFX-PheX-3_at	3662.97095	1700.94287	2160.51758	1693.8429	2224.62549	1571.963[illegible]	[illegible]
AFFX-ThrX-5_at	10185.57129	5153.46045	6371.82617	5580.12646	5664.89648	5009.4995[illegible]	[illegible]11.46777
AFFX-ThrX-M_at	10491.83789	4495.55127	4837.53613	4179.62109	4377.70361	3976.51733	4847.24072
AFFX-ThrX-3_at	11368.03711	5048.31396	5490.34375	5003.65918	4917.28809	4614.06299	5325.36572
AFFX-TrpnX-5_at	9.33082	29.28475	18.24754	13.6274	9.38387	11.135	11.39099
AFFX-TrpnX-M_at	6.59241	24.26511	11.11075	9.2284	11.96512	10.81012	12.87228
AFFX-TrpnX-3_at	4.19613	0.00495	5.6787	7.44433	3.21669	7.53645	12.73567
AFFX-r2-Ec-bioB-5_at	144.67793	314.88318	162.42134	112.69901	107.30542	174.22957	106.59195
AFFX-r2-Ec-bioB-M_at	198.60356	270.05942	114.38917	84.17824	92.68085	143.71248	95.73489
AFFX-r2-Ec-bioB-3_at	113.14938	113.97368	77.19824	93.50596	81.25646	109.75388	97.33247
AFFX-r2-Ec-bioC-5_at	140.88416	105.04796	70.25456	84.09558	68.96983	104.08804	83.34672
AFFX-r2-Ec-bioC-3_at	162.81339	75.68188	65.35989	54.97689	54.26827	76.12438	64.88766
AFFX-r2-Ec-bioD-5_at	473.12143	470.50519	483.51208	246.47073	258.4718	362.00604	306.15591
AFFX-r2-Ec-bioD-3_at	97.41333	64.63611	82.89006	78.40298	63.10776	99.59647	86.99286
AFFX-r2-P1-cre-5_at	573.66705	2.96001	10.51656	3.10805	458.20541	553.59656	7.17838
AFFX-r2-P1-cre-3_at	675.42224	0.01487	5.39451	0.01057	535.53625	816.06439	1.5461

Here is some data I got for mouse liver exposed to phenobarbitol, please analyze it for me. Thanks, your collaborator -Y.S.

Statistical and scientific databases aim to organize quantitative data and information in ways that incorporate existing understanding and lead to new understanding by creating a model of the system.

- Observation and Mechanism (both what and why)
  - Description ← → explanation within a context
  - [GCCGCCGCC] ← → fragile X syndrome





* Scientific and Statistical Databases Management Conference Proceedings from 2011 http://ssdbm2011.ssdbm.org/sessionnotes.php#

What is a Database?

- In the most generic terms a database is an organized collection of information.

- Why does it need to be organized?

What is database design?

- Database design at its most basic level is the logical manner in which the data/information is organized and stored.

- So, why is database design important?
  - What are the basic design considerations be?

Basic Database Design

- To be able to engineer a solution you must first:

"Know your data!"

Frank Crane ~ "You may be deceived if you trust too much, but you will live in torment if you don't trust enough."

Database design considerations

- What is the data that is going to be stored?
- What is the goal of the system?
- How much data?
  - Do you have MB or TB?
- How big are the pieces of information?
  - How much data do you have entering the database?
  - Being retrieved from the database?
  - What form is the data in when you go to put it in the database?
    - Is the form consistent? (*date fields*)
      - If not how do you handle variation?
  - Are all the data types consistent and the same?
  - Do the data points conflict? *(SOF in crystal structures)*
    - Does your definition cover the bases? (*bytea*)
  - Is the information unique or are their replicated entries?