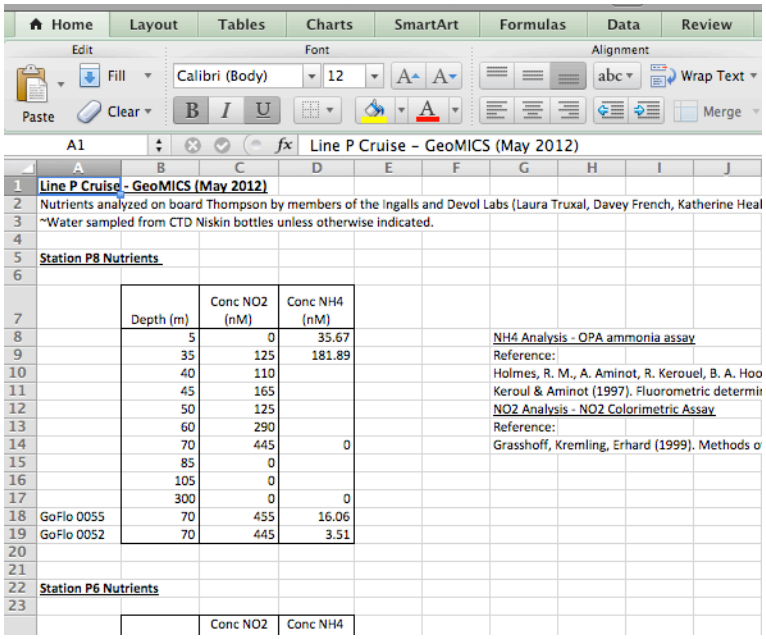
UNIVERSITY of WASHINGTON

Data Models and Databases

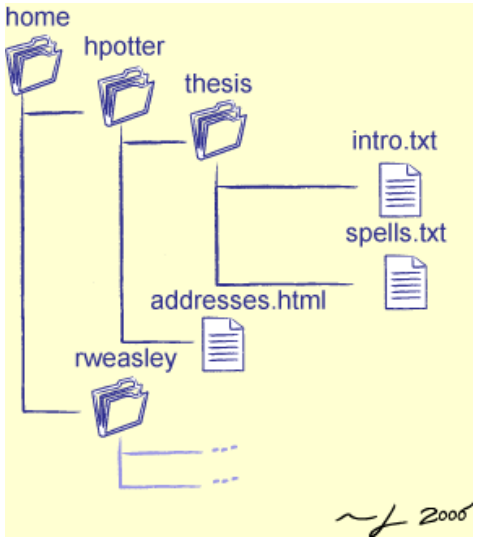
Bill Howe, PhD
Director of Research,
Scalable Data Analytics
University of Washington
eScience Institute

How do we store data?

How do we store data?



	A	B	C	D	E	F	G	H	I	J
1	Line P Cruise - GeoMICS (May 2012)									
2	Nutrients analyzed on board Thompson by members of the Ingalls and Devol Labs (Laura Truxal, Davey French, Katherine Heal									
3	~Water sampled from CTD Niskin bottles unless otherwise indicated.									
4										
5	Station P8 Nutrients									
6										
7		Depth (m)	Conc NO2 (nM)	Conc NH4 (nM)						
8		5	0	35.67			NH4 Analysis - OPA ammonia assay			
9		35	125	181.89			Reference:			
10		40	110				Holmes, R. M., A. Aminot, R. Kerouel, B. A. Hoo			
11		45	165				Keroul & Aminot (1997). Fluorometric determi			
12		50	125				NO2 Analysis - NO2 Colorimetric Assay			
13		60	290				Reference:			
14		70	445	0			Grasshoff, Kremling, Erhard (1999). Methods o			
15		85	0							
16		105	0							
17		300	0	0						
18	GoFlo 0055	70	455	16.06						
19	GoFlo 0052	70	445	3.51						
20										
21										
22	Station P6 Nutrients									
23										
			Conc NO2	Conc NH4						



ANNOTATIONSUMMARY-COMBINEDORFANNOTATION16_Phaeo_genome

###query	length	COG hit #1	e-value #1	identity #1	score #1	hit length #1	description #1
chr_4[480001-580000].287	4500						
chr_4[560001-660000].1	3556						
chr_9[400001-500000].503	4211	COG4547	2.00E-04	19	44.6	620	Cobalamin biosynthesis proteir
chr_9[320001-420000].548	2833	COG5406	2.00E-04	38	43.9	1001	Nucleosome binding factor SPN
chr_27[320001-404298].20	3991	COG4547	5.00E-05	18	46.2	620	Cobalamin biosynthesis proteir
chr_26[320001-420000].378	3963	COG5099	5.00E-05	17	46.2	777	RNA-binding protein of the Puf
chr_26[400001-441226].196	2949	COG5099	2.00E-04	17	43.9	777	RNA-binding protein of the Puf
chr_24[160001-260000].65	3542						
chr_5[720001-820000].339	3141	COG5099	4.00E-09	20	59.3	777	RNA-binding protein of the Puf
chr_9[160001-260000].243	3002	COG5077	1.00E-25	26	114	1089	Ubiquitin carboxyl-terminal hyc
chr_12[720001-820000].86	2895	COG5032	2.00E-09	30	60.5	2105	Phosphatidylinositol kinase and
chr_12[800001-900000].109	1463	COG5032	1.00E-09	30	60.1	2105	Phosphatidylinositol kinase and
chr_11[1-100000].70	2886						
chr_11[80001-180000].100	1523						

What is the *data model*?

What is a Data Model?

Three components:

1. Structures
2. Constraints
3. Operations

Examples

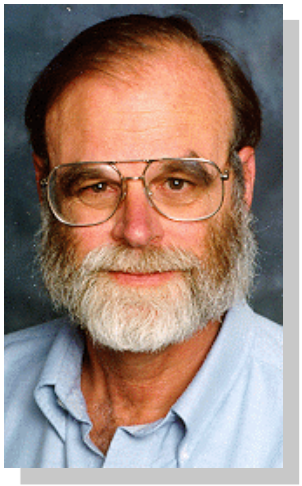
1. Structures
   - rows and columns?
   - nodes and edges?
   - key-value pairs?
   - a sequence of bytes?
2. Constraints
   - all rows must have the same number of columns
   - all values in one column must have the same type
   - a child cannot have two parents
3. Operations
   - find the value of key x
   - find the rows where column “lastname” is “Jordan”
   - get the next N bytes

What is a database?

A collection of information organized to afford efficient retrieval

http://www.usg.edu/galileo/skills/unit04/primer04_01.phtml

Another view



"When people use the word database, fundamentally what they are saying is that the data should be self-describing and it should have a schema. That's really all the word database means."

-- Jim Gray, "The Fourth Paradigm"

Why would I want a database?

What problem do they solve?

1. Sharing
   Support concurrent access by multiple readers and writers
2. Data Model Enforcement
   Make sure all applications see clean, organized data
3. Scale
   Work with datasets too large to fit in memory
4. Flexibility
   Use the data in new, unanticipated ways

Questions to consider

- How is the data physically organized on disk?
- What kinds of queries are efficiently supported by this organization, and what kinds are not?
- How hard is it update the data, or add new data?
- What happens when I encounter new queries that I didn't anticipate? Do I reorganize the data? How hard is that?