Data Formats and Tables

CMSE 890-602

What is data?

What is data?

- Text
- Tables
- Time series

- Images - tables of color information
- Videos - time series of images
- Sound - time series of frequencies
- etc

Computational data: Text vs Binary

- Text is stored as characters
  - Easy to version control
  - Easy to read
  - Potentially human readable
  - Large file sizes (but can be compressed as binary data)
- Binary data is stored as a direct binary representation
  - Not human readable
  - Difficult to version control
  - Potentially smaller file sizes

Compression

- Converts data to a binary representation
- Reduces file size by reducing duplication of information
- More redundant data = better compression
- Typically data must be decompressed in its entirety to be read
- Some compressed formats can be decompressed into memory for access

Basic data: Tables

- Straightforward storage of related data
- Columns define the content
- Rows store the content
- Index accesses rows

Index	Column 1	Column 2	Column 3
0	Data	Data	
1		Data	
2	Data		

Common table operations

- Index a row or selection of rows
- Select a column or selection of columns
- Compute summary statistics
  - Row or column direction
- Create new rows or columns
- Transpose the table (swap rows and columns)

Index	Column 1	Column 2	Column 3	
0	Data	Data		Summary row 0
1		Data		
2	Data			
		Summary Column 2		

Human vs Machine-readable Tables

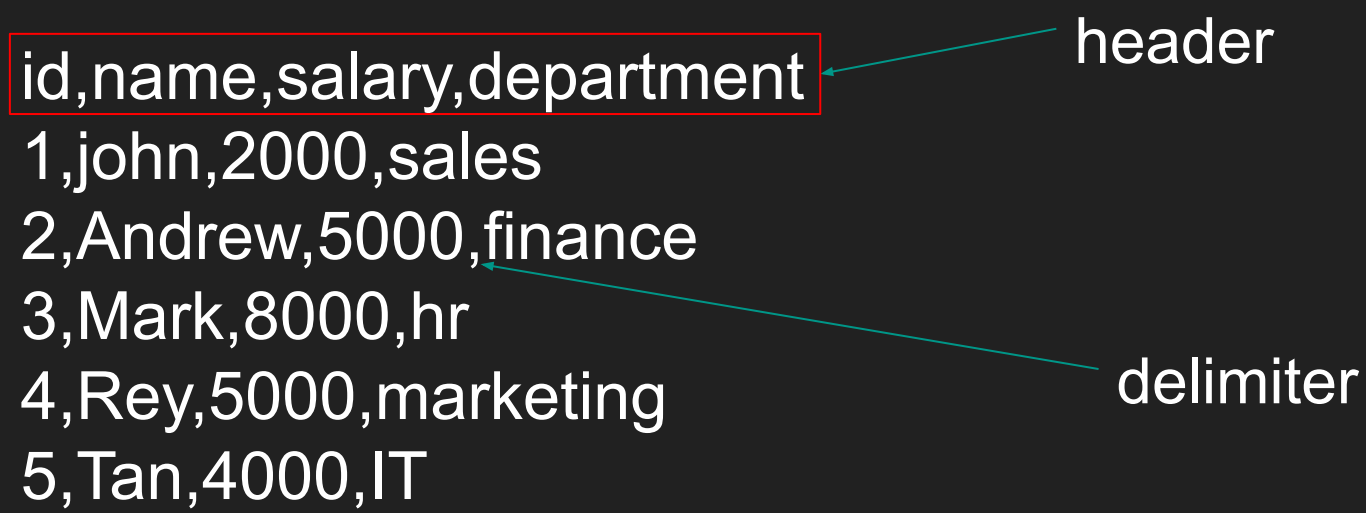
- Humans typically prefer whitespace or lines as a delimiter for tables
- Computers work best with specific characters e.g. ,
- Special characters in text (especially from non-English languages) may be difficult to handle
- Parsing (reading) tables using software may require custom solutions
    - But many solutions exist e.g. Pandas has read_csv()

A bad but human-readable format

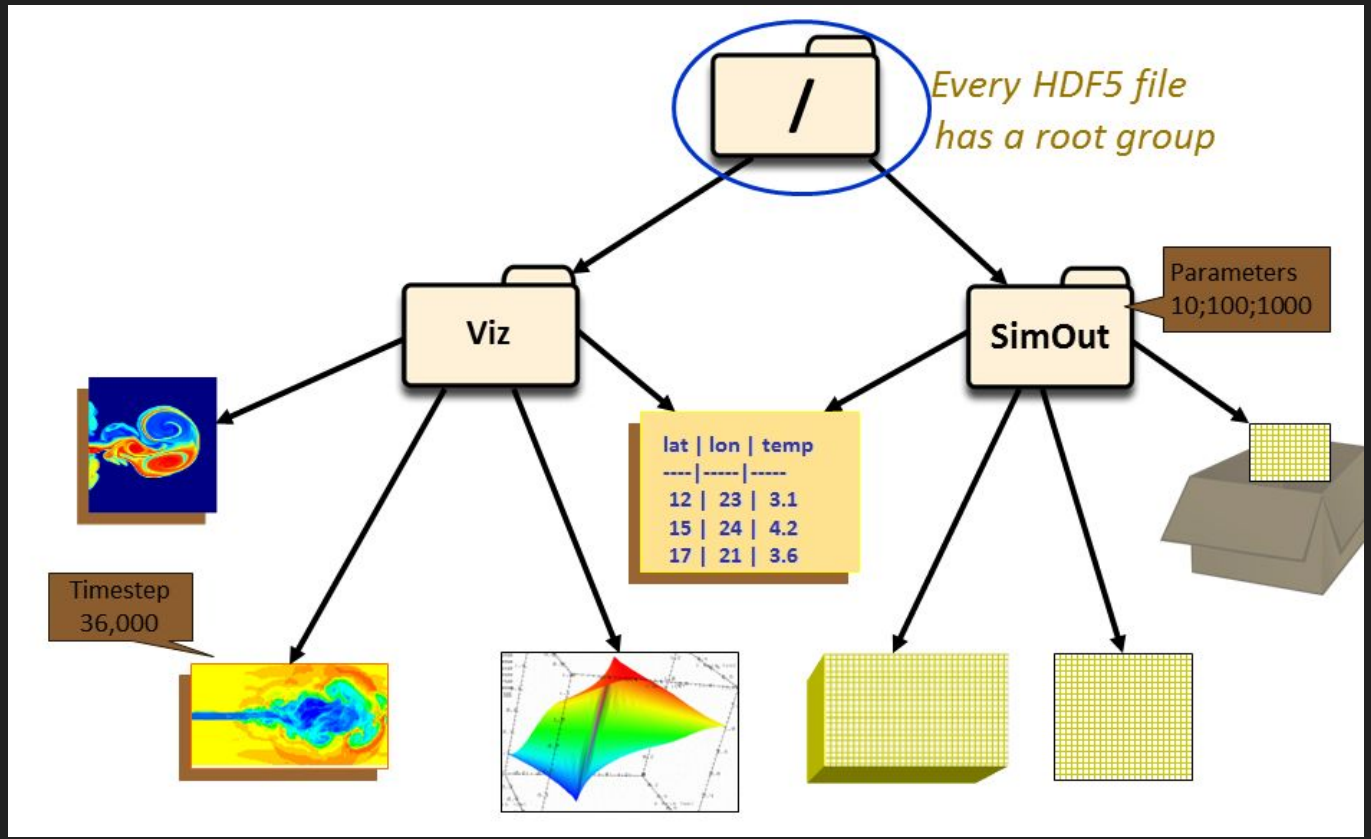
WR134 - time sequence polarimetry (Crimea 1989, values plus errors)

JD-2447700	U				B				V				R				I			
	q		u		q		u		q		u		q		u					
59.3018	1.204	.048	0.358	.079	0.946	.032	0.218	.048	1.091	.062	0.303	.068	0.839	.028	0.190	.059	0.777	.023	0.155	.053
60.4033	1.051	.034	0.265	.025	0.872	.039	0.179	.028	0.912	.040	0.203	.026	0.717	.021	0.114	.039	0.621	.037	0.168	.021
60.4692	0.946	.070	0.395	.060	0.790	.050	0.229	.054	0.858	.046	0.268	.068	0.738	.064	0.215	.045	0.622	.048	0.204	.043
61.2734	1.182	.059	0.441	.104	0.990	.034	0.256	.077	1.015	.039	0.275	.081	0.793	.029	0.245	.072	0.739	.045	0.175	.053
61.3306	1.360	.108	0.340	.089	1.145	.044	0.244	.065	1.261	.051	0.176	.071	0.920	.081	0.167	.054	0.859	.050	0.115	.056
61.4160	1.284	.082	0.468	.086	1.051	.032	0.331	.041	1.161	.060	0.398	.057	0.995	.077	0.200	.073	0.897	.050	0.225	.056
61.5273	1.500	.097	0.340	.107	1.009	.047	0.277	.019	1.148	.041	0.292	.034	0.960	.062	0.214	.059	0.889	.075	0.176	.061
62.3887	1.106	.063	0.191	.053	0.978	.034	0.070	.031	1.035	.039	0.162	.058	0.783	.033	0.033	.039	0.689	.018	0.002	.033

CSV format


id,name,salary,department
1,john,2000,sales
2,Andrew,5000,finance
3,Mark,8000,hr
4,Rey,5000,marketing
5,Tan,4000,IT
header
delimiter

HDF: a common binary format

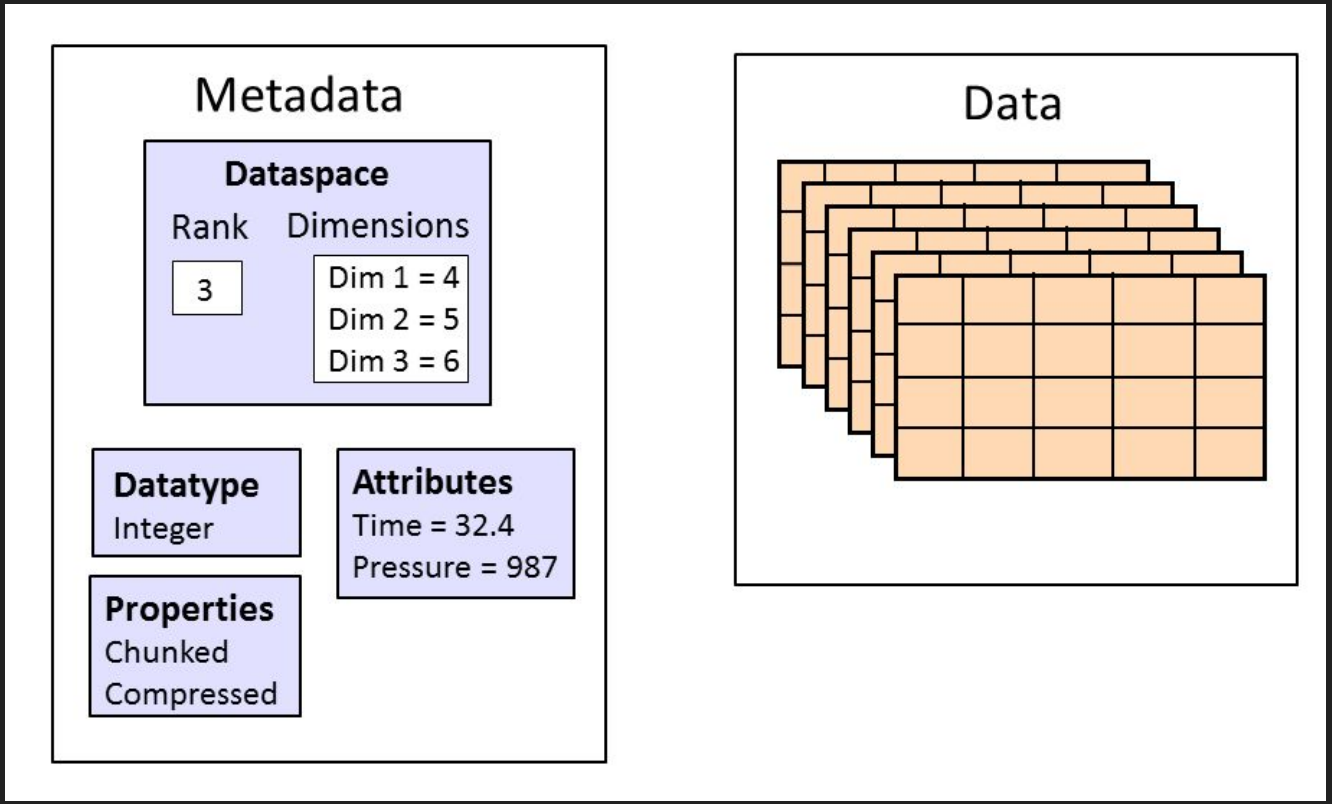


https://hdfgroup.github.io/hdf5/_intro_h_d_f5.html

Metadata: Data about data

- Critical for reproducibility
  - Should be included with **any** release of data!
- Provides information about:
  - Data types
  - Author
  - Source (physical or computational)
  - License for use
- Most basic form would be a human-readable text file with the data file

Metadata example with HDF



https://hdfgroup.github.io/hdf5/_intro_h_d_f5.html

FAIR data

- Findability
- Accessibility
- Interoperability
- Reusability

https://www.go-fair.org/fair-principles/

Findability

- Data is searchable and can be searched for
- This requires *metadata* (data about data!)
  - Labels
  - References
  - Provenance (where did it come from?)
  - Uniqueness
- Data should be stored in a location that can be searched in some way
- Usually the metadata is what can be searched for

Accessibility

- The data can be read by both humans and machines
- The metadata includes clear format information
- Examples of how to read the data may be provided
- Metadata can always be read even if the data is gone

Interoperability

- The data works with multiple machine types
  - Operating systems
  - Programming languages
- References other data where necessary

Reusability

- Data are clearly described
- Data are clearly licensed
- Relevant community standards are met
  - These are often different between communities!

Storage locations for data

- Data Dryad https://datadryad.org/
- Zenodo https://zenodo.org/
- ACCESS long-term storage https://allocations.access-ci.org/resources
- MSU digital commons https://works.hcommons.org/
- All are searchable and handle metadata control
- Some include automatic DOI minting

Activity

- On D2L access the in-class assignment “In-class 8: Data formats” and follow the instructions
- PDF of instructions will be posted to the class repository as well

Pre-class 9: Databases

- Go to https://www.w3schools.com/sql/ and learn about SQL
- Try to complete the first 12 exercises at https://www.w3schools.com/sql/exercise.asp
- Post a screenshot of your results on D2L