Data Types, Sources and Tasks

Alark Joshi

Slides credits: T. Moeller

Basic Variable Types

- Physical Type
  - Characterized by storage format and machine operations
  - Example: bool, short, int, float, double, string,…
- Abstract type
  - Provide descriptions of the data
  - Characterized by attributes
  - May be organized into a hierarchy
  - Example: cars, bicycles, …

Data Values

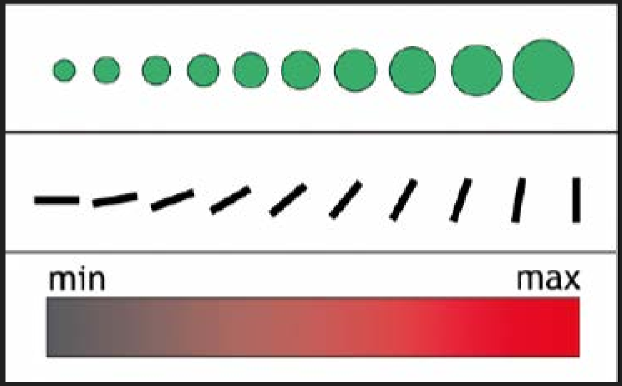
- Characteristics of data values
  - Range of values
  - Data types
  - Quantitative data types (scalar, vector, tensor; kind of discretization)
  - Dimension (number of components)
  - Error (variance)

Data Values

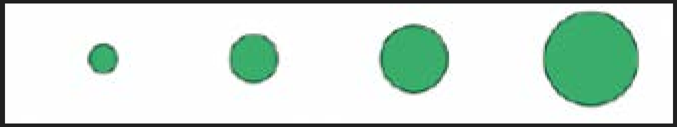
- Range of values
  - Qualitative
    - Non-metric
    - Ordinal (order along a scale)
    - Nominal (no order)
  - Quantitative
    - Metric scale
    - Discrete
    - Continuous
    - Interval/ratio

Data Types

- Quantitative (Q)
  - 10 inches, 20 inches, etc.



- Ordinal (ordered) (O)
  - Small, Medium, Large



- Nominal (categorical) (N)
  - Apples, Oranges, Bananas, …

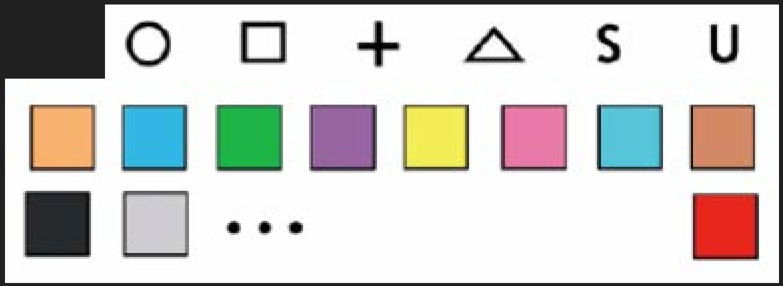


Image credits: Stolte et al, Polaris

Quantitative

- Q – Interval (location of zero is arbitrary)
  - Dates: Jan 31st; Location in terms of Lat, Long
  - Only differences (i.e. intervals) can be compared

- Q – Ratio (zero fixed)
  - Measurements: Length, Mass, Temperature, …
  - Origin is meaningful and helps with comparison tasks (ratios, proportions)

Quantitative

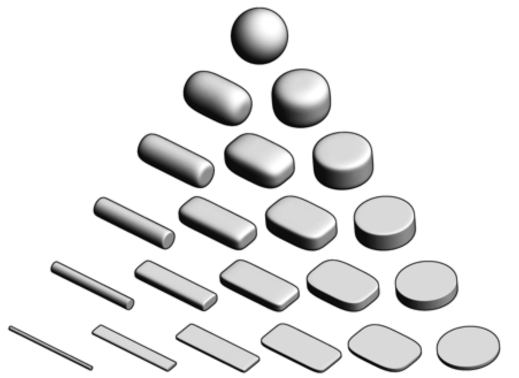


Image credits: G. Kindlmann

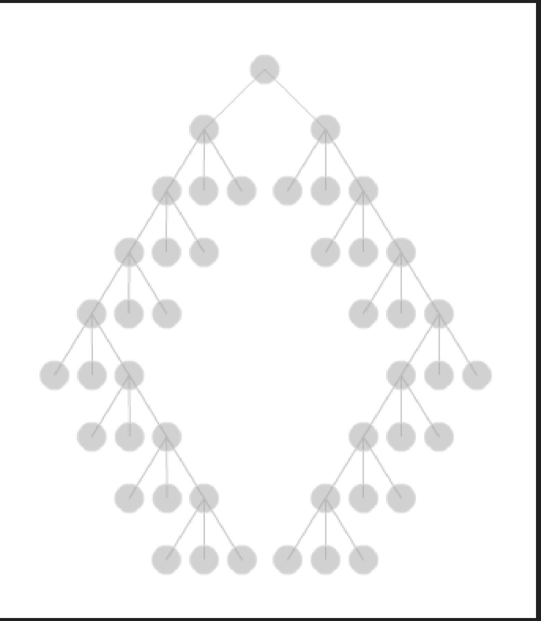
- Scalar Data is given by a function $f(x_1, \ldots, x_n): R^n \rightarrow R$ with n independent variables x_i
- Vector Data represents direction and magnitude and is given by an m-tuple $(f_1, \ldots f_m)$ with $f_k = f_k(x_1, \ldots, x_n)$, $m \geq 2$ and $1 \leq k \leq m$
  - Usually m = n unless there is projection involved
- Tensor data for a tensor of level k is given by $t_{i1,i2,\ldots,ik}(x_1, \ldots, x_n)$
  - A tensor of level 1 is a vector, a tensor of level 2 is a matrix

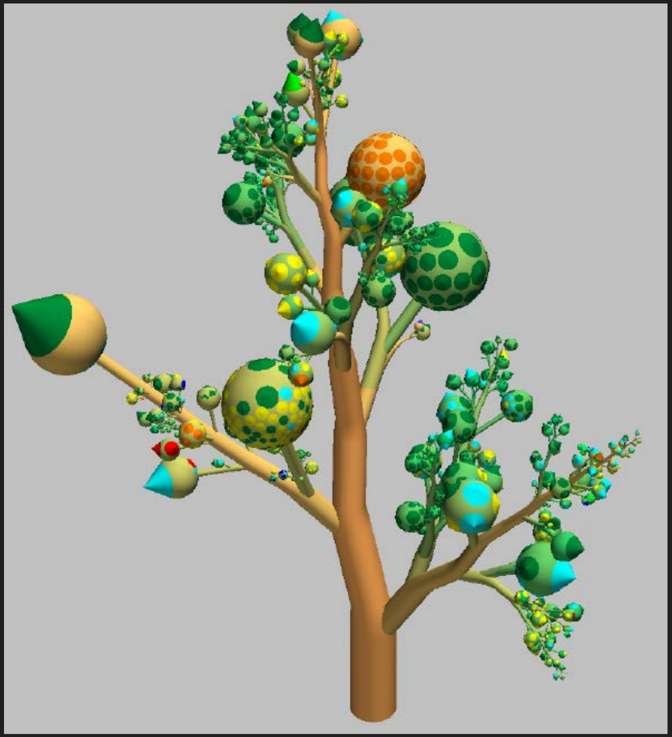
Quantitative - Time

- Discretization in time with constant or variable time steps
- Time dependency of
  - Data only (grid remains constant)
    - E.g. Time series of CT data, CFD simulation of air around the wing of an airplane
  - Data and grid geometry (topology remains constant)
    - E.g. Crash testing of a car
  - Data, grid geometry and topology
    - E.g. engine simulation with a moving piston

Structure of the data

- Sequential (list)
- Relation (table)
- Hierarchical (tree)
- Network structure (graph)





A	B	C	S	T	U
Order ID	Order Date	Order Priority	Product Container	Product Base Margin	Ship Date
3	10/14/06	5-Low	Large Box	0.8	10/21/06
6	2/21/08	4-Not Specified	Small Pack	0.55	2/22/08
32	7/16/07	2-High	Small Pack	0.79	7/17/07
32	7/16/07	2-High	Jumbo Box	0.72	7/17/07
32	7/16/07	2-High	Medium Box	0.6	7/18/07
32	7/16/07	2-High	Medium Box	0.65	7/18/07
35	10/23/07	4-Not Specified	Wrap Bag	0.52	10/24/07
35	10/23/07	4-Not Specified	Small Box	0.58	10/25/07
36	11/3/07	1-Urgent	Small Box	0.55	11/3/07
65	3/18/07	1-Urgent	Small Pack	0.49	3/19/07
66	1/20/05	5-Low	Wrap Bag	0.56	1/20/05
69	6/4/05	4-Not Specified	Small Pack	0.44	6/6/05
69	6/4/05	4-Not Specified	Wrap Bag	0.6	6/6/05
70	12/18/06	5-Low	Small Box	0.59	12/23/06
70	12/18/06	5-Low	Wrap Bag	0.82	12/23/06
96	4/17/05	2-High	Small Box	0.55	4/19/05
97	1/29/06	3-Medium	Small Box	0.38	1/30/06
129	11/19/08	5-Low	Small Box	0.37	11/28/08
130	5/8/08	2-High	Small Box	0.37	5/9/08
130	5/8/08	2-High	Medium Box	0.38	5/10/08
130	5/8/08	2-High	Small Box	0.6	5/11/08
132	6/11/06	3-Medium	Medium Box	0.6	6/12/06
132	6/11/06	3-Medium	Jumbo Box	0.69	6/14/06
134	5/1/08	4-Not Specified	Large Box	0.82	5/3/08
135	10/21/07	4-Not Specified	Small Pack	0.64	10/23/07
166	9/12/07	2-High	Small Box	0.55	9/14/07
193	8/8/06	1-Urgent	Medium Box	0.57	8/10/06
194	4/5/08	3-Medium	Wrap Bag	0.42	4/7/08

Record

A	B	C	S	T	U
Order ID	Order Date	Order Priority	Product Container	Product Base Margin	Ship Date
3	10/14/06	5-Low	Large Box	0.8	10/21/06
6	2/21/08	4-Not Specified	Small Pack	0.55	2/22/08
32	7/16/07	2-High	Small Pack	0.79	7/17/07
32	7/16/07	2-High	Jumbo Box	0.72	7/17/07
32	7/16/07	2-High	Medium Box	0.6	7/18/07
32	7/16/07	2-High	Medium Box	0.65	7/18/07
35	10/23/07	4-Not Specified	Wrap Bag	0.52	10/24/07
35	10/23/07	4-Not Specified	Small Box	0.58	10/25/07
36	11/3/07	1-Urgent	Small Box	0.55	11/3/07
65	3/18/07	1-Urgent	Small Pack	0.49	3/19/07
66	1/20/05	5-Low	Wrap Bag	0.56	1/20/05
69	6/4/05	4-Not Specified	Small Pack	0.44	6/6/05
69	6/4/05	4-Not Specified	Wrap Bag	0.6	6/6/05
70	12/18/06	5-Low	Small Box	0.59	12/23/06
70	12/18/06	5-Low	Wrap Bag	0.82	12/23/06
96	4/17/05	2-High	Small Box	0.55	4/19/05
97	1/29/06	3-Medium	Small Box	0.38	1/30/06
129	11/19/08	5-Low	Small Box	0.37	11/28/08
130	5/8/08	2-High	Small Box	0.37	5/9/08
130	5/8/08	2-High	Medium Box	0.38	5/10/08
130	5/8/08	2-High	Small Box	0.6	5/11/08
132	6/11/06	3-Medium	Medium Box	0.6	6/12/06
132	6/11/06	3-Medium	Jumbo Box	0.69	6/14/06
134	5/1/08	4-Not Specified	Large Box	0.82	5/3/08
135	10/21/07	4-Not Specified	Small Pack	0.64	10/23/07
166	9/12/07	2-High	Small Box	0.55	9/14/07
193	8/8/06	1-Urgent	Medium Box	0.57	8/10/06
194	4/5/08	3-Medium	Wrap Bag	0.42	4/7/08

Field / Dimension

A	B	C	S	T	U
Order ID	Order Date	Order Priority	Product Container	Product Base Margin	Ship Date
3	10/14/06	5-Low	Large Box	0.8	10/21/06
6	2/21/08	4-Not Specified	Small Pack	0.55	2/22/08
32	7/16/07	2-High	Small Pack	0.79	7/17/07
32	7/16/07	2-High	Jumbo Box	0.72	7/17/07
32	7/16/07	2-High	Medium Box	0.6	7/18/07
32	7/16/07	2-High	Medium Box	0.65	7/18/07
35	10/23/07	4-Not Specified	Wrap Bag	0.52	10/24/07
35	10/23/07	4-Not Specified	Small Box	0.58	10/25/07
36	11/3/07	1-Urgent	Small Box	0.55	11/3/07
65	3/18/07	1-Urgent	Small Pack	0.49	3/19/07
66	1/20/05	5-Low	Wrap Bag	0.56	1/20/05
69	6/4/05	4-Not Specified	Small Pack	0.44	6/6/05
69	6/4/05	4-Not Speci	[illegible]	0.6	6/6/05
70	12/18/06	5-Low	[illegible]	0.59	12/23/06
70	12/18/06	5-Low	[illegible]	0.82	12/23/06
96	4/17/05	2-High	[illegible]	0.55	4/19/05
97	1/29/06	3-Medium	[illegible]	0.38	1/30/06
129	11/19/08	5-Low	Small Box	0.37	11/28/08
130	5/8/08	2-High	Small Box	0.37	5/9/08
130	5/8/08	2-High	Medium Box	0.38	5/10/08
130	5/8/08	2-High	Small Box	0.6	5/11/08
132	6/11/06	3-Medium	Medium Box	0.6	6/12/06
132	6/11/06	3-Medium	Jumbo Box	0.69	6/14/06
134	5/1/08	4-Not Specified	Large Box	0.82	5/3/08
135	10/21/07	4-Not Specified	Small Pack	0.64	10/23/07
166	9/12/07	2-High	Small Box	0.55	9/14/07
193	8/8/06	1-Urgent	Medium Box	0.57	8/10/06
194	4/5/08	3-Medium	Wrap Bag	0.42	4/7/08

1 = Quantitative

2 = Nominal

3 = Ordinal

A	B	C	S	T	U
Order ID	Order Date	Order Priority	Product Container	Product Base Margin	Ship Date
3	10/14/06	5-Low	Large Box	0.8	10/21/06
6	2/21/08	4-Not Specified	Small Pack	0.55	2/22/08
32	7/16/07	2-High	Small Pack	0.79	7/17/07
32	7/16/07	2-High	Jumbo Box	0.72	7/17/07
32	7/16/07	2-High	Medium Box	0.6	7/18/07
32	7/16/07	2-High	Medium Box	0.65	7/18/07
35	10/23/07	4-Not Specified	Wrap Bag	0.52	10/24/07
35	10/23/07	4-Not Specified	Small Box	0.58	10/25/07
36	11/3/07	1-Urgent	Small Box	0.55	11/3/07
65	3/18/07	1-Urgent	Small Pack	0.49	3/19/07
66	1/20/05	5-Low	Wrap Bag	0.56	1/20/05
69	6/4/05	4-Not Specified	Small Pack	0.44	6/6/05
69	6/4/05	4-Not Speci		0.6	6/6/05
70	12/18/06	5-Low		0.59	12/23/06
70	12/18/06	5-Low		0.82	12/23/06
96	4/17/05	2-High		0.55	4/19/05
97	1/29/06	3-Medium		0.38	1/30/06
129	11/19/08	5-Low	Small Box	0.37	11/28/08
130	5/8/08	2-High	Small Box	0.37	5/9/08
130	5/8/08	2-High	Medium Box	0.38	5/10/08
130	5/8/08	2-High	Small Box	0.6	5/11/08
132	6/11/06	3-Medium	Medium Box	0.6	6/12/06
132	6/11/06	3-Medium	Jumbo Box	0.69	6/14/06
134	5/1/08	4-Not Specified	Large Box	0.82	5/3/08
135	10/21/07	4-Not Specified	Small Pack	0.64	10/23/07
166	9/12/07	2-High	Small Box	0.55	9/14/07
193	8/8/06	1-Urgent	Medium Box	0.57	8/10/06
194	4/5/08	3-Medium	Wrap Bag	0.42	4/7/08
194	4/5/08	3-Medium	Wrap Bag	0.84	4/7/08

1 = Quantitative

2 = Nominal

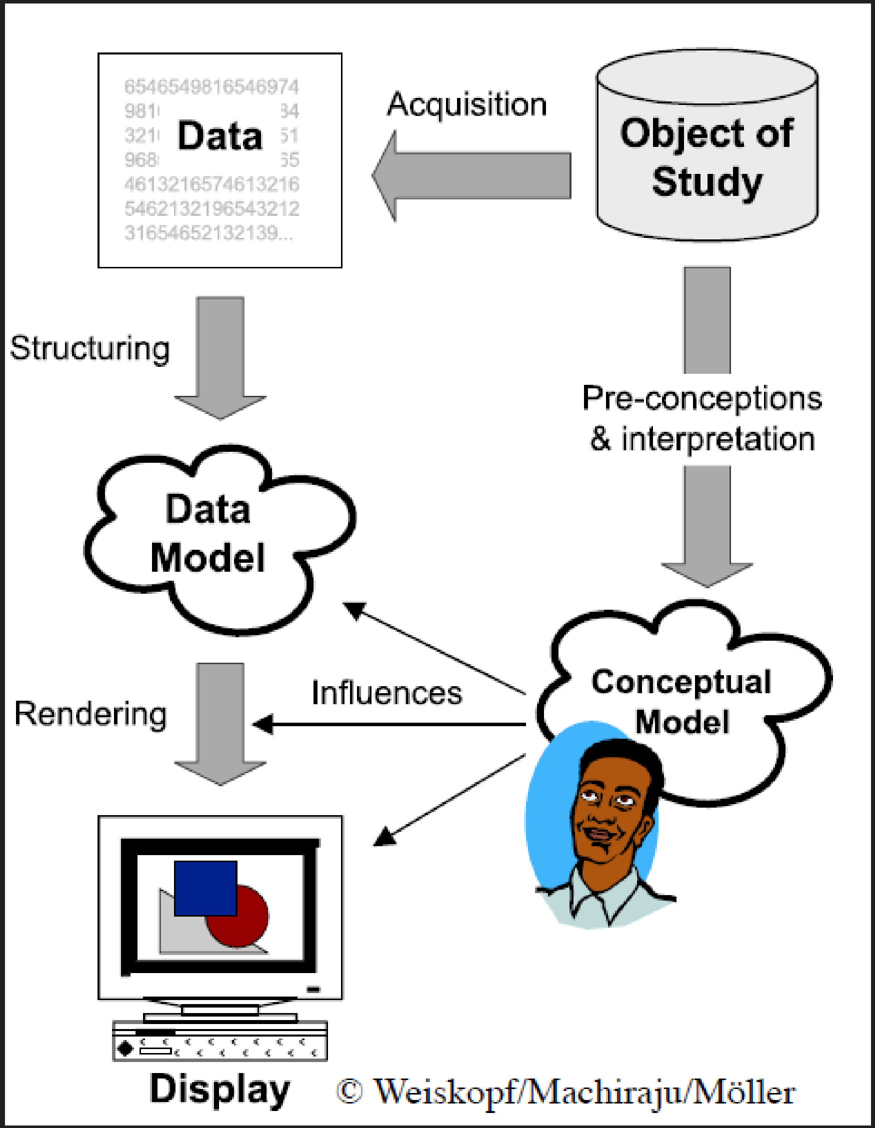
3 = Ordinal

A	B	C	S	T	U
Order ID	Order Date	Order Priority	Product Container	Product Base Margin	Ship Date
3	10/14/06	5-Low	Large Box	0.8	10/21/06
6	2/21/08	4-Not Specified	Small Pack	0.55	2/22/08
32	7/16/07	2-High	Small Pack	0.79	7/17/07
32	7/16/07	2-High	Jumbo Box	0.72	7/17/07
32	7/16/07	2-High	Medium Box	0.6	7/18/07
32	7/16/07	2-High	Medium Box	0.65	7/18/07
35	10/23/07	4-Not Specified	Wrap Bag	0.52	10/24/07
35	10/23/07	4-Not Specified	Small Box	0.58	10/25/07
36	11/3/07	1-Urgent	Small Box	0.55	11/3/07
65	3/18/07	1-Urgent	Small Pack	0.49	3/19/07
66	1/20/05	5-Low	Wrap Bag	0.56	1/20/05
69	6/4/05	4-Not Specified	Small Pack	0.44	6/6/05
69	6/4/05	4-Not Specified	Wrap Bag	0.6	6/6/05
70	12/18/06			0.59	12/23/06
70	12/18/06			0.82	12/23/06
96	4/17/05			0.55	4/19/05
97	1/29/06	3-Medium	Small Box	0.38	1/30/06
129	11/19/08	5-Low	Small Box	0.37	11/28/08
130	5/8/08	2-High	Small Box	0.37	5/9/08
130	5/8/08	2-High	Medium Box	0.38	5/10/08
130	5/8/08	2-High	Small Box	0.6	5/11/08
132	6/11/06	3-Medium	Medium Box	0.6	6/12/06
132	6/11/06	3-Medium	Jumbo Box	0.69	6/14/06
134	5/1/08	4-Not Specified	Large Box	0.82	5/3/08
135	10/21/07	4-Not Specified	Small Pack	0.64	10/23/07
166	9/12/07	2-High	Small Box	0.55	9/14/07
193	8/8/06	1-Urgent	Medium Box	0.57	8/10/06
194	4/5/08	3-Medium	Wrap Bag	0.42	4/7/08
194	4/5/08	3-Medium	Wrap Bag	0.84	4/7/08

Quantitative = Measures

Nominal /Ordinal = Dimensions

Data vs. Conceptual Models



Data vs. Conceptual Models

- Data model – low level description of data
  - Set with operations such as floats with +, -, /, *
- Conceptual Model requires mental refining
  - Includes semantics, supports reasoning

Data	Conceptual
1D floats	temperature
3D vector of floats	space

Example

- From data model – 32.5, 54.0, … (floats)
- Using conceptual model (temperature)
- To data type
  - Continuous to 4 significant figures (Q)
  - Hot, warm, cold (O)
  - Burned vs. not burned (N)

Dimension

- Number of variables per class
  - 1 = Univariate
  - 2 = Bivariate
  - 3 = Trivariate
  - >3 = Hypervariate or Multidimensional or Multiattribute

Data Tasks

- "Information exploration should be a joyous experience" – Shneiderman
- Example: Baby Name Voyager
- Visualizations provide context to enable selection and regions and provide dynamic feedback
- Shneiderman proposes a type by task taxonomy for information visualization

B. Shneiderman, The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations, Visual Languages, 1996.

Task Abstraction

- Overview: Get an overview of the entire data
- Zoom: Zoom in on items of interest
- Filter: Filter out uninteresting items
- Details on demand: Select an item or group and get details when needed

B. Shneiderman, The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations, Visual Languages, 1996.

Task Abstraction

- Related: View relationships among items
- History: Keep a history of actions to support undo, replay, and progressive refinement
- Extract: Allow extraction of sub-collections and of query parameters

Shneiderman’s Mantra

- Visual Information Seeking Mantra
- Overview first, zoom and filter, then details-on-demand

Examples

- 1-dimensional dataset –
  - Bifocal display
  - http://bost.ocks.org/mike/fisheye/

- Filmfinder – Developed queries and sliders for user controlled visualization of multidimensional data (http://hcil.cs.umd.edu/video/1994/1994_visualinfo.mpg)

Task Abstraction

- Amar, Egan and Stasko's study led to an analytic task taxonomy
- Ten tasks from the analysis are

1. Retrieve Value
2. Filter
3. Compute Derived Value
4. Find Extremum
5. Sort
6. Determine Range
7. Characterize Distribution
8. Find Anomalies
9. Cluster
10. Correlate

Task Abstraction

- Example Task:
  - What is the car with the highest MPG?
  - What director has won the most awards?
  - 4. Find Extremum will help you answer these kinds of questions easily
- Example Task: Determine the average calorie content of Post cereals?
  - 3. Compute Derived Value will let you run queries and create new derived values to perform the task

Task Abstraction

- Example: [Amar, Eagan and Stasko, 2005]
  - Order the cars by weight
  - Rank the cereal by calories

1. Filter: Find data that satisfies the conditions
2. Find Extremum: Find data with extreme values
3. Sort: Rank data according to metric of choice
4. Determine Range: Find span of data values
5. Find anomalies: Find data with unexpected/extreme values

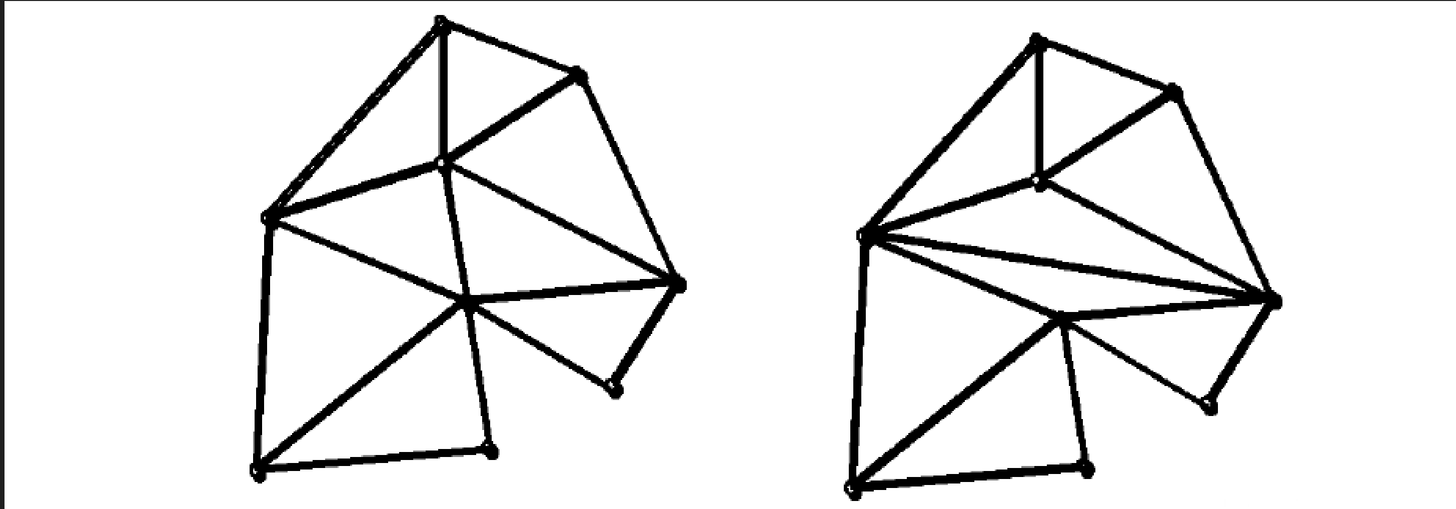
Data Grids

Topology

- If points are arbitrarily distributed and there is no connectivity between them, the data is called scattered
- Otherwise, data is composed of cells bounded by grid lines
- Topology specifies the connectivity of data
- Geometry specifies the position of the data

Topology

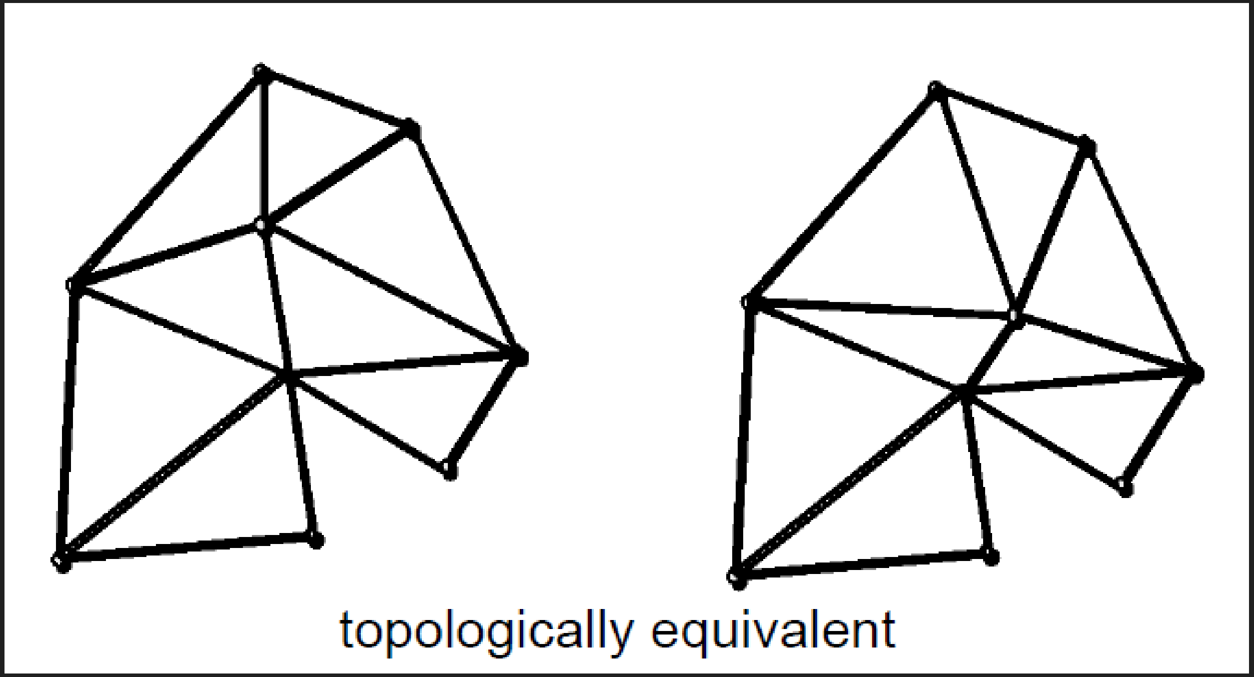
- Properties of geometric shapes that remain unchanged even when under distortion



Same geometry (vertex positions), different topology (connectivity)

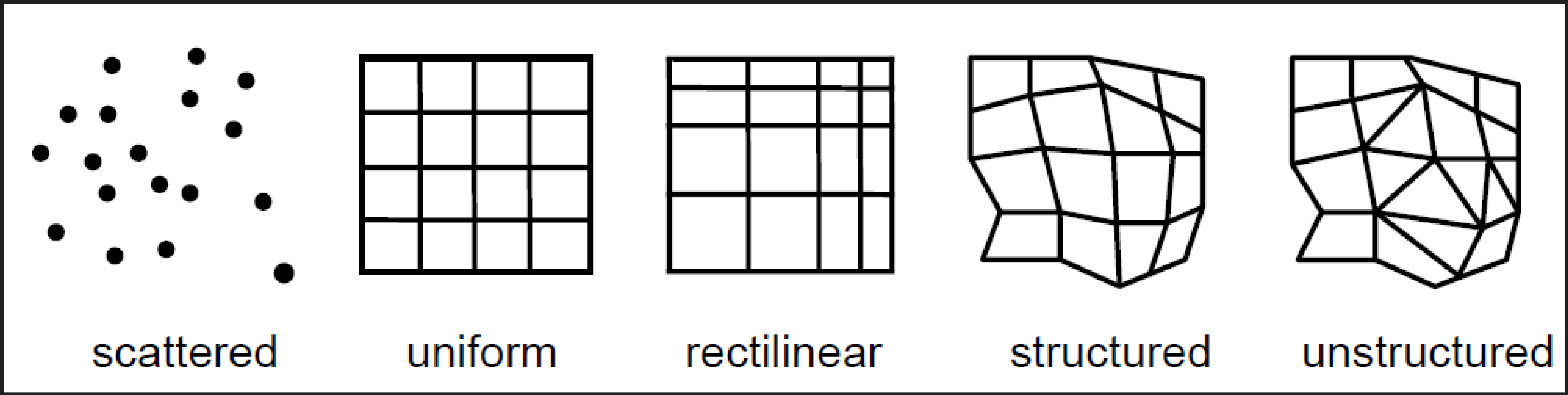
Topologically equivalent

- Things that can be transformed into each other by stretching and squeezing, without tearing or sticking together bits which were previously separated



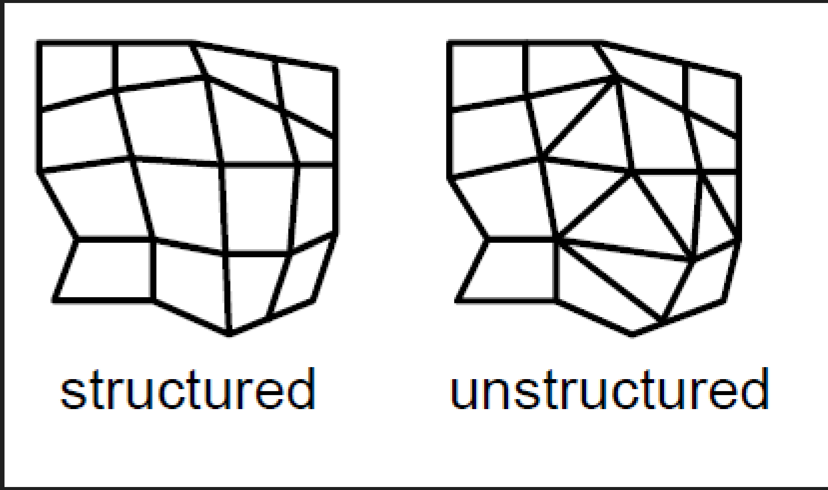
Grid Types

- Grid differ substantially in the cells (building blocks) they are constructed from and in the way the topological information is specified

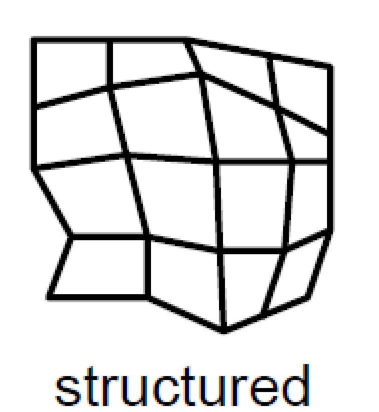


Structured and Unstructured Grids

- Structured grids have a regular topology and regular/irregular geometry
- Unstructured grids have irregular topology and geometry



Characteristics of Structured Grids



- Easier to compute with
- May require more elements or unevenly shaped elements to precisely cover the underlying domain
- Topology is represented implicitly by an n-vector of dimensions
- Geometry is represented explicitly by an array of points
- Every interior point has the same number of neighbors

Characteristics of Unstructured Grids

- If no implicit topological information is given, the grids are called unstructured grids
- Grid point geometry and connectivity must be stored
- Dedicated data structures efficient traversal and data
- Often composed of triangl
- Typically, fewer elements domain

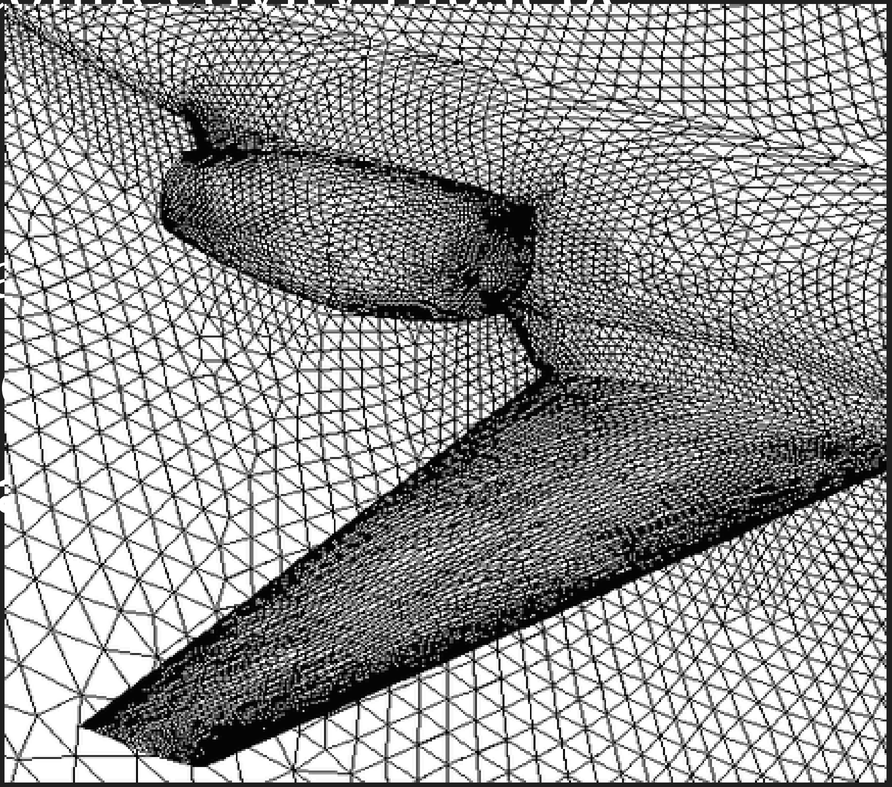
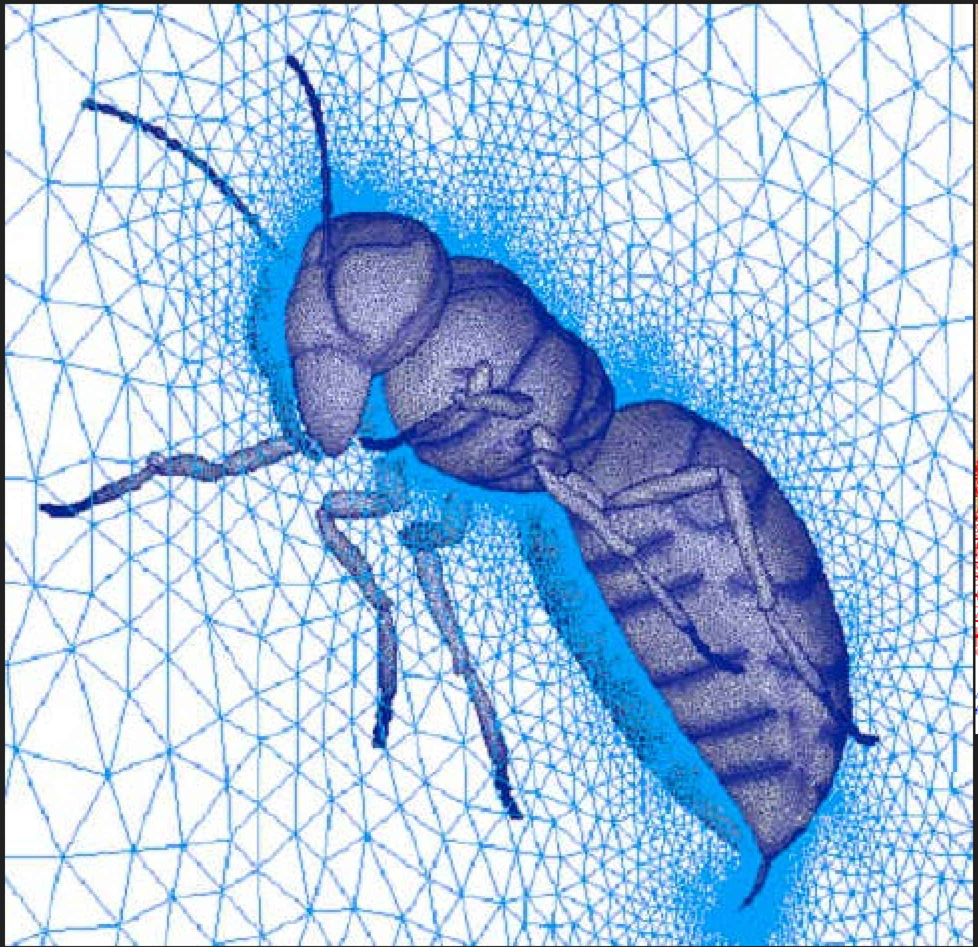
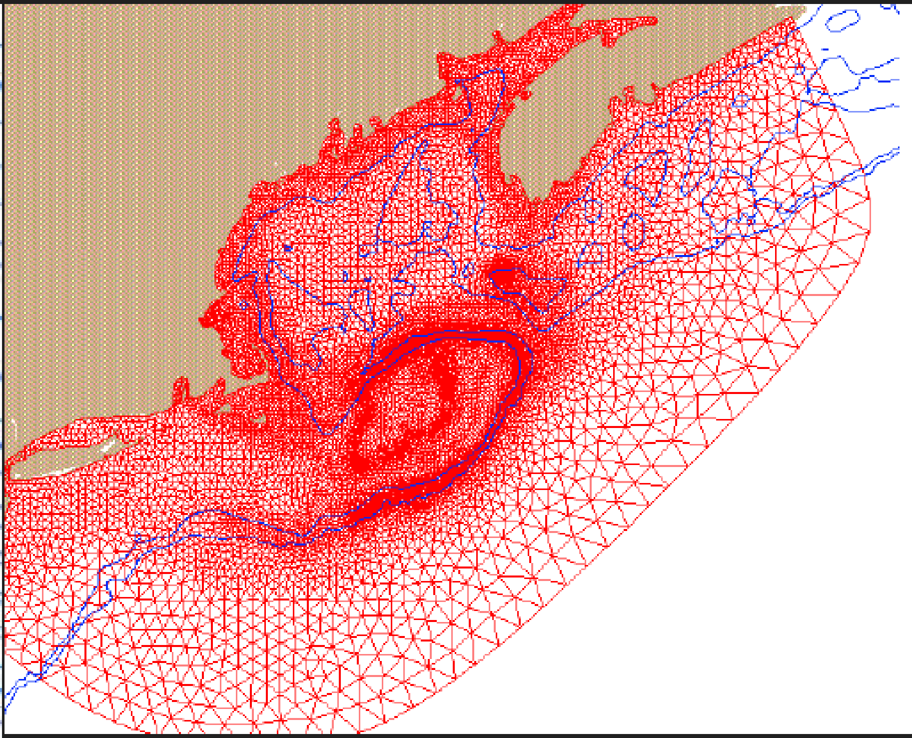


Image credits: http://geolab.larc.nasa.gov/GridTool/Training/VGRID/

Unstructured Grids

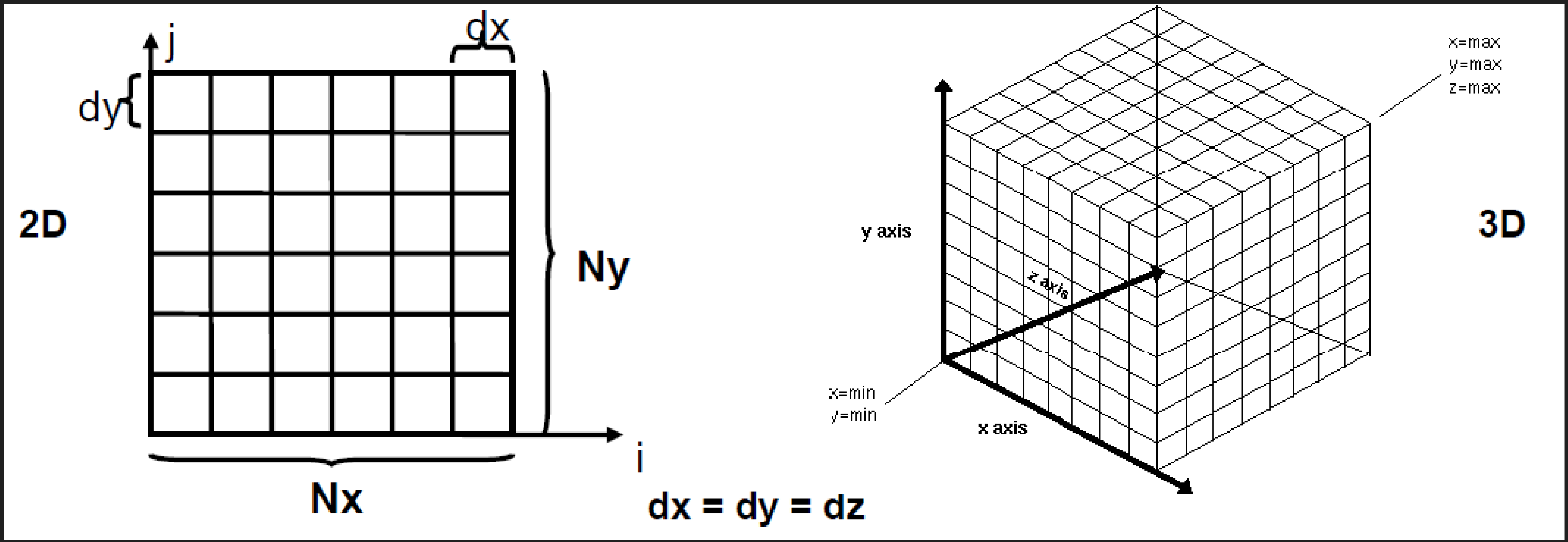
- Can be adapted to capture local features





Types of grids

- Cartesian or equidistant grids
  - Structured grid
  - Number of points = Nx * Ny * Nz

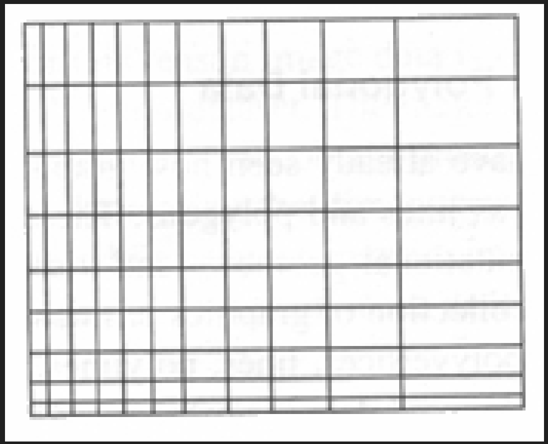
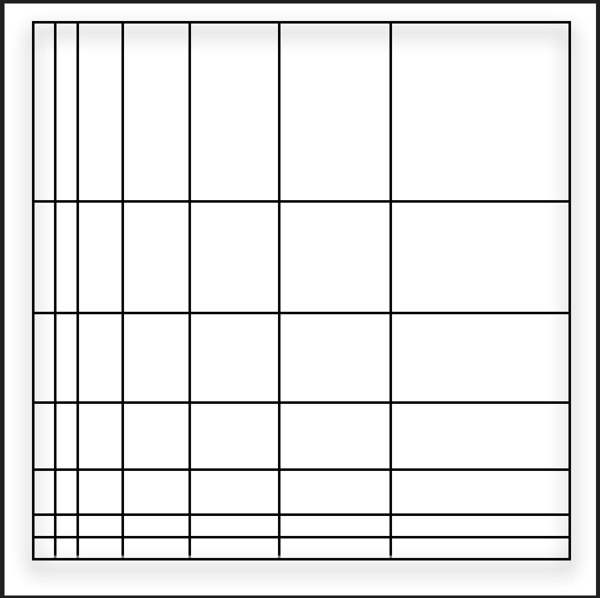


Types of grids

- Uniform grids are similar to Cartesian grids
- Consist of equal cells but with different resolution in at least one dimension ($dx \neq dy \neq dz$)
- Typical example is medical imaging data that consists of slices
  - Slice images with square pixels ($dx = dy$)
  - Larger slice distance ($dz > dx = dy$)

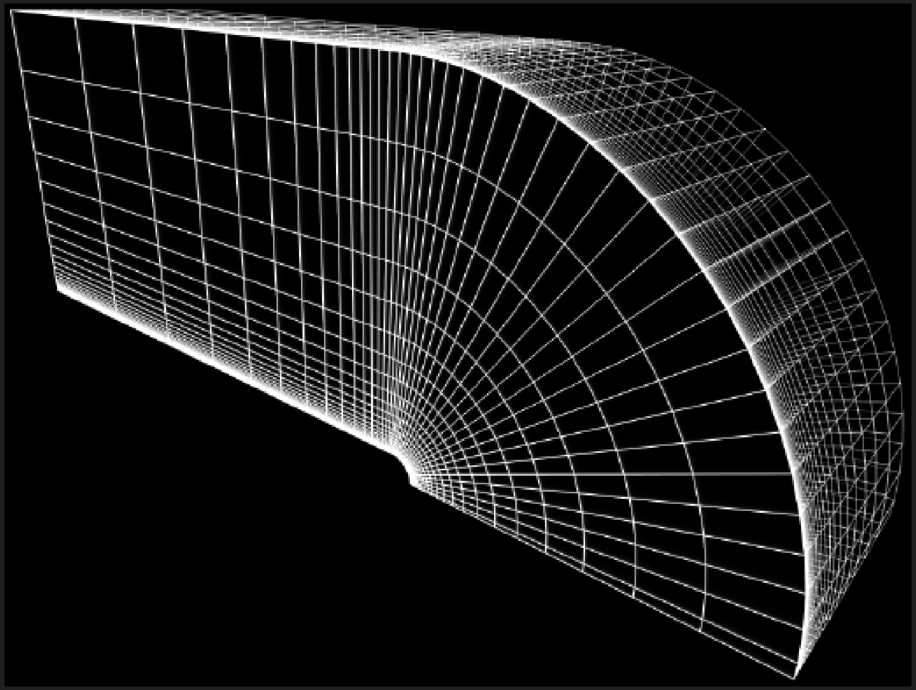
Types of grids

- Rectilinear grids
- Topology is still regular but irregular spacing between grid points
- Topology is still implicit

Types of grids

- Curvilinear grids
  - Topology is still regular but irregular spacing between grid points
  - Topology is implicit, but vertex positions are explicitly stored



Types of grids

- Multigrids
  - Focus in specific area to avoid unnecessary detail in other areas
- Finer grid for regions of interest
- Difficulties at the boundaries between low and high res grids for operations such as interpolation