EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Data Sources and Representations

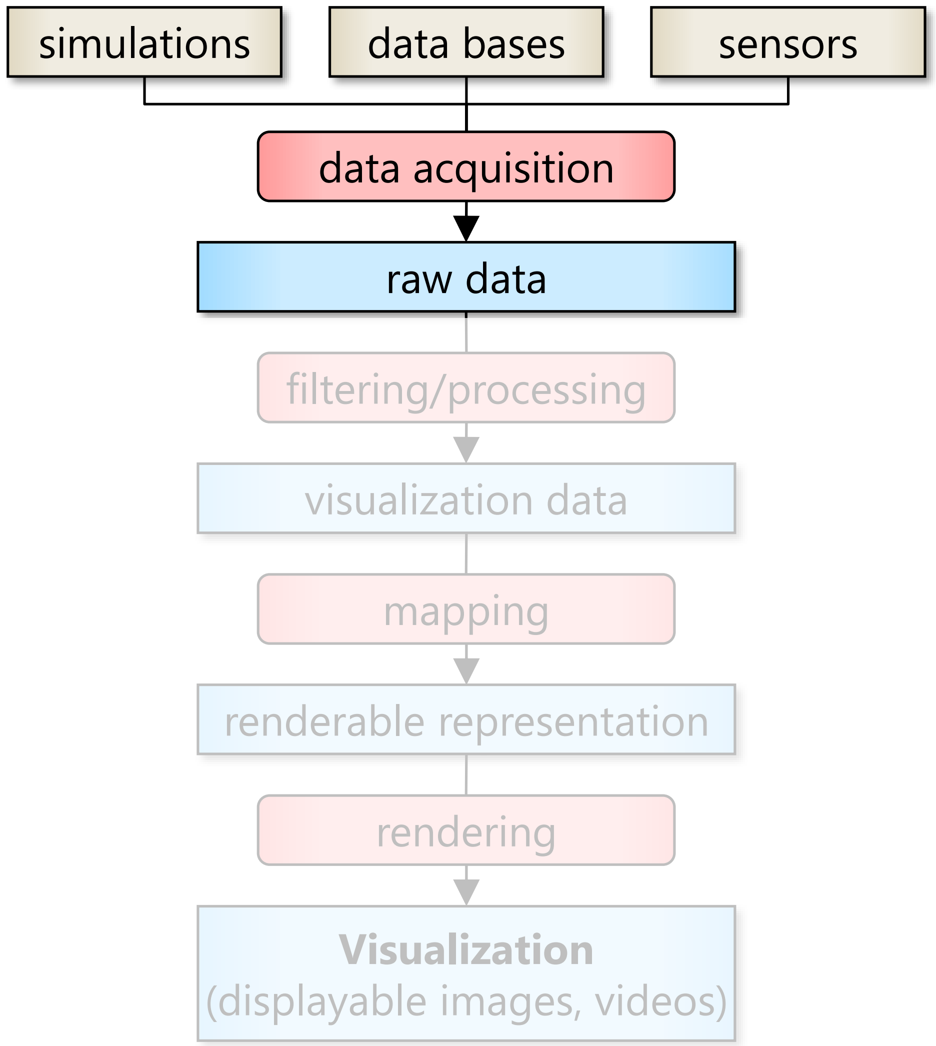
Scientific Visualization – Summer Semester 2021

Jun.-Prof. Dr. **Michael Krone**

Contents

- Data sources
  - Data acquisition with scanners
  - Sources of error
- Data representation
  - Domain
  - Data structures
  - Data values
  - Data classification

Focus:
First part of visualization pipeline



Data Sources

- The capability of traditional presentation techniques is not sufficient for the increasing amount of data to be interpreted
  - Data might come from any source with almost arbitrary size
  - Techniques to efficiently visualize large-scale data sets and new data types need to be developed
- Real world
  - Measurements and observation
- Theoretical world
  - Mathematical and technical models
- Artificial world
  - Data that is designed

Data Sources – **Real-world (Measurements)**

- Medical Imaging (MRI, CT, PET)
- Geographical information systems (GIS)
- Electron microscopy

GB

- Meteorology and environmental sciences (satellites)
- Seismic data
- Crystallography

TB

- High energy physics
- Astronomy (e.g. Solar Dynamics Observatory 1.5 TB/day)
- Defense

PB

Data Sources – Theoretical world

- Computer simulations
  - Sciences
    - Molecular dynamics
    - Quantum chemistry
    - Mathematics

    GB

    - Molecular modeling
    - Computational physics
    - Meteorology
    - Computational fluid mechanics (CFD)

    TB
  - Engineering
    - Architectural walk-throughs

    GB

    - Structural mechanics
    - Car body design

    TB

Data Sources – **Theoretical world**

- Computer simulations
  - Sciences
    - Molecular dynamics
    - Quantum chemistry
    - Mathematics

    GB

    - Molecular modeling
    - Computational physics
    - Meteorology
    - Computational fluid mechanics (CFD)

    TB
  - Engineering
    - Architectural walk-throughs

    GB

    - Structural mechanics
    - Car body design

    TB

Data Sources – **Theoretical World**

- Computer simulations
  - Commercial
    - Business graphics **MB**
    - Economic models **TB**
    - Financial modeling
- Information systems **PB**
  - Stock market (300 million transactions per day in NY)
  - Market and sales analysis
  - World Wide Web

Data Sources – **Artificial World**

- Drawings
- Painting
- Publishing

MB

- TV (teasers, commercials)

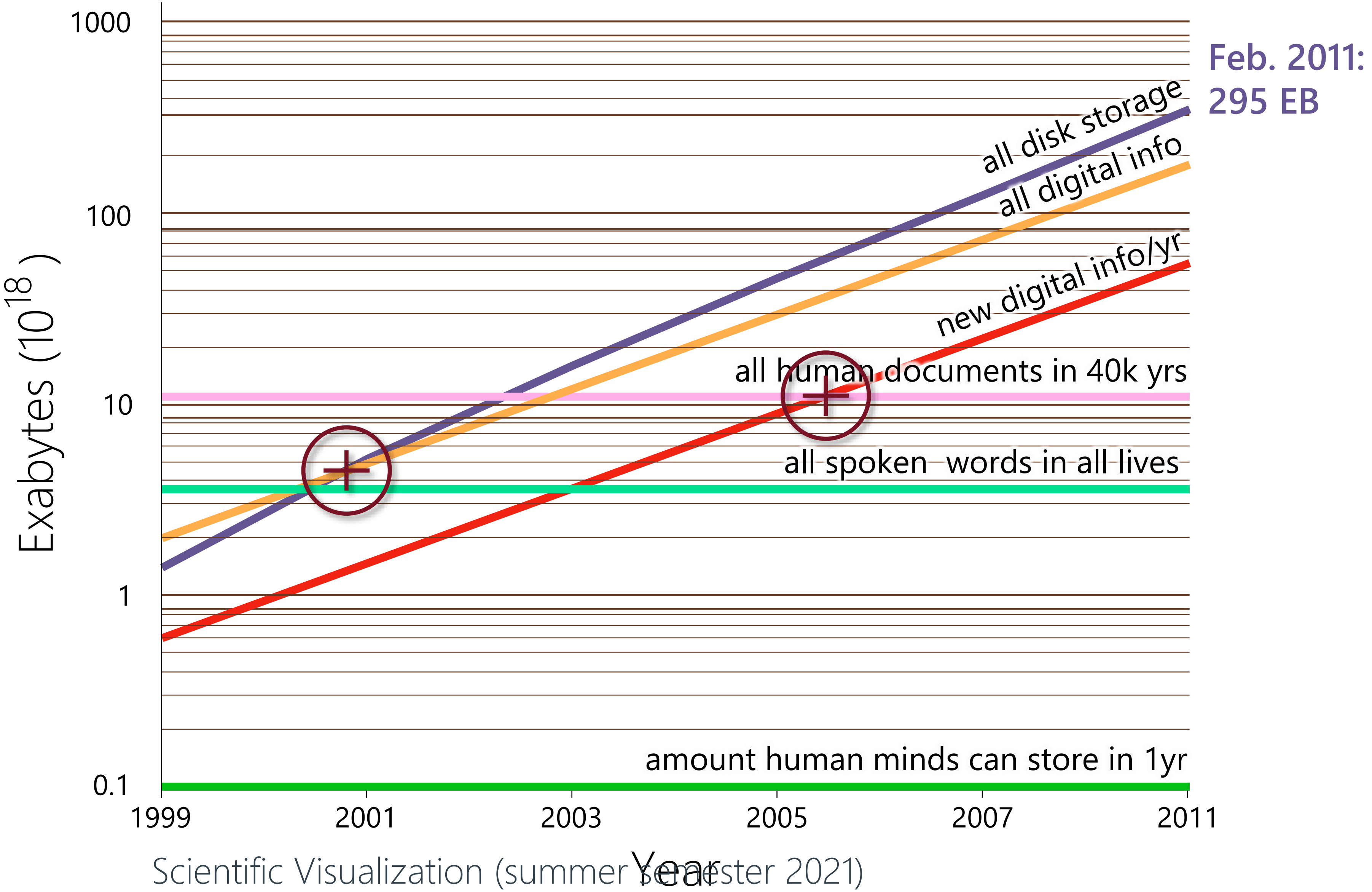
GB

- Movies (animations, special effects)

TB

Data Sources – Information Explosion

- Every two days we create as much data as we did from the beginning of mankind until 2003!

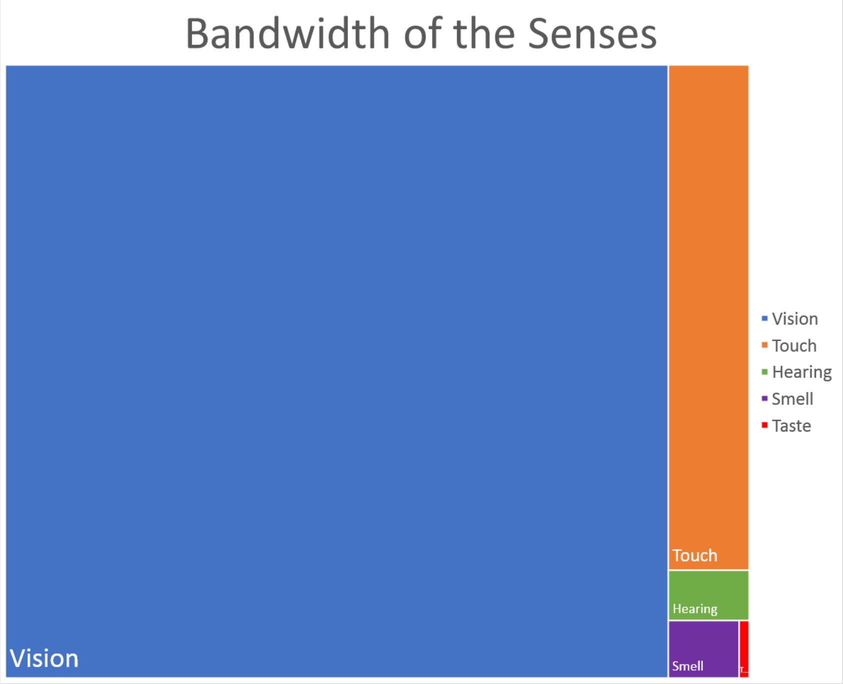


Sources: Lesk, Berkeley SIMS, Landauer, EMC, TechCrunch, Smart Planet (Image courtesy of C. Johnson, SCI / University of Utah)

Data Sources – Information Explosion

- Today at home: terabyte disks, gigaflops CPUs, gigabit Ethernet
- Todays high-end: petabyte RAIDs, teraflops GPUs
- Todays peak: the peta era (10^{15})
  - HPC: petaflop/s → *Summit* supercomputer at ORNL
  - Sensors: petabytes/year → Large Hadron Collider at CERN: 15 PB per year
  - Data: Google processes ≥25 PB per day
  - Digital Events: peta-events/year → IP packets DE-CIX backbone
- Tomorrow: the era of exa (10^{18}), zetta (10^{21}), yotta (10^{24}),…
- Digital Information (created, captured, replicated)
  - Since 2010 almost 1 zettabyte increase per year (28T e-mails/yr, ~6 EB of data)
  - Only 5% structured information (text, numbers, …)

Big Data – Visualization



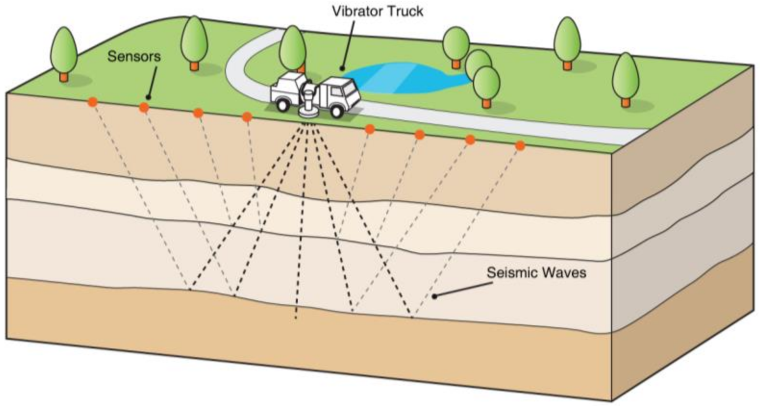
- Information “stored” and processed by humans
  - est. some petabytes in entire life time
  - 80% through vision (space, form, color, texture,…)
  - Visual cortex and related functions occupy about half of our brains
  - Vision is the highest bandwidth human-computer interface
    - Total bandwidth of human sight: ~10,000,000 Bits/second
    - Looking at displays with 1-100 Megapixels
    - Data rate per person (US study): 30 GB per day (≥1% text, 50% interactively)

- **Visualization** plays a significant role in dealing with digital data
- **Interaction** and **abstraction** are key for the visualization of huge data

Data Acquisition with Scanners

- Medical scanners:
  - X-rays
  - Computed Tomography (CT)
  - MRI (or NMR)
  - PET / SPECT
  - Ultrasound
- Other examples:
  - PIV (particle image velocimetry): experimental flow measurement
  - X-rays/MRT for material science
  - Seismic data (oil and gas industry)

Particle Image Velocimetry (www.dantecdynamics.com)



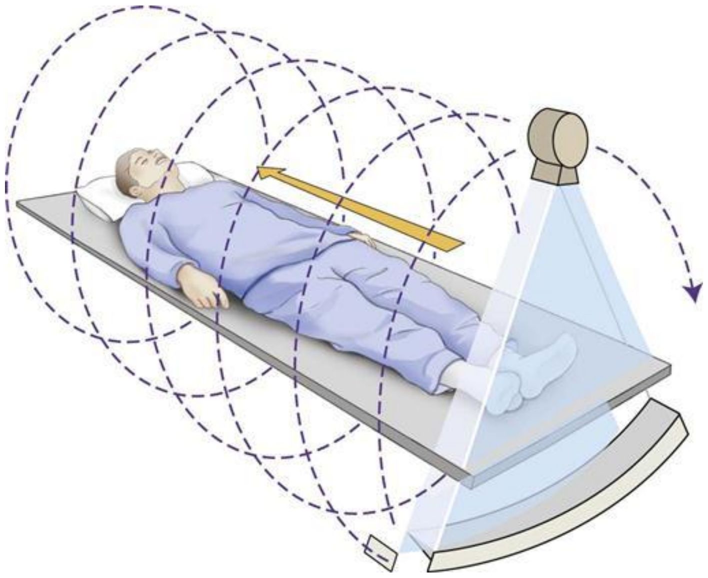
Seismic (http://www.innoseis.com/seismic-surveying/)

Data Acquisition with Scanners

- **X-ray**
  - Bones contain heavy atoms
    - Act as an absorber of X-rays
  - Commonly used to image bone structure and lungs
  - Excellent for detecting metal objects
  - Main disadvantage:
    - Lack of complete anatomical structure
    - All other tissue has very similar absorption coefficient for X-rays
    - Soft tissue X-ray absorption relatively high → health risk

Data Acquisition with Scanners

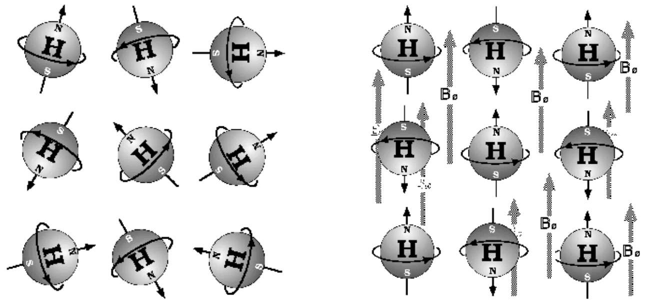
- **Computed** (Axial) **Tomography** (CT)
  - Improves traditional X-ray imaging
  - Based on the principle that a 3D object can be reconstructed from its 2D projections
  - Combine X-ray images from various angles
  - Advantages
    - Superior to single X-ray scans
    - Easier to separate soft tissues (materials other than bone)
    - Data exist in digital form: can be analyzed quantitatively
  - Disadvantages
    - Significantly more data collected
    - Soft tissue X-ray absorption still similarly high → health risk



Data Acquisition with Scanners

- **Nuclear Magnetic Resonance** (NMR) or **Magnetic Resonance Imaging** (MRI)
  - Polarization through external magnetic field
  - A second magnetic field is applied to excite nuclear spins
  - Measure: radiation from relaxation
  - 3D position from gradients in second magnetic field
  - MRI is especially sensitive for hydrogen (H)
    - Can measure diffusion of water molecules

    → Diffusion tensor imaging (see later)
  - Advantages:
    - Detailed anatomical information
    - No high-energy radiation, i.e. “safe” scanning method

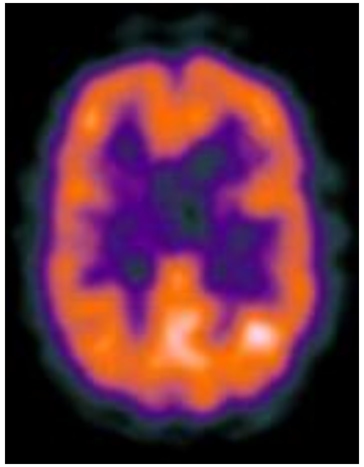


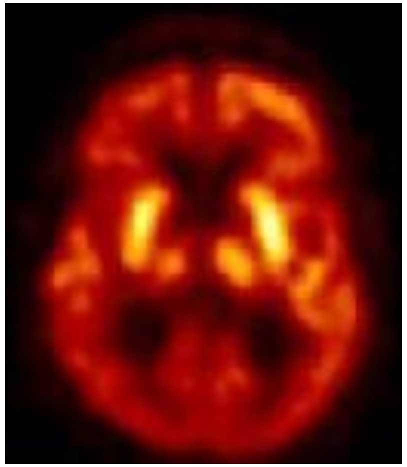
Data Acquisition with Scanners

- **Positron Emission Tomography** (PET) and **Single Photon Emission Computerized Tomography** (SPECT)
  - Scans the emission of particles by compounds injected into the body
  - Follow the movements of the injected compound and its metabolism
  - Reconstruction techniques similar to CT

SPECT

- Emit (any) gamma rays
- Collected with gamma camera





PET

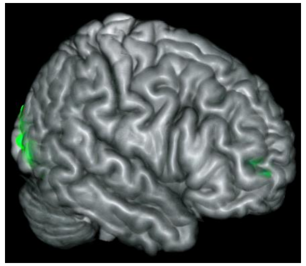
- Positron collides with electron to emit photons in 180° angle
- Both annihilation photons detected in coincidence
- Higher sensitivity

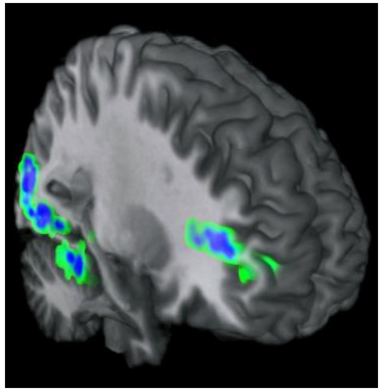
Data Acquisition with Scanners

- **Ultrasound**
  - High-frequency sound (ultrasonic) waves
    - Above the range of sound audible to humans (typically >1 MHz)
    - Piezoelectric crystal creates sound waves
  - Change in tissue density reflects waves
  - Echoes are recorded
  - Delay of reflected signal and amplitude determines the position of the tissue
  - Properties
    - Noise-affected
    - 1D, 2D, 3D scanners
    - Irregular sampling – reconstruction problems

Sources of Error

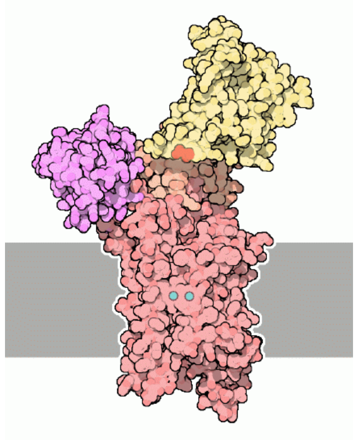
- Data acquisition
  - Accuracy and reliability of scanner?
  - Sampling: are we sampling data with enough precision to get what we need?
  - Quantization: are we converting "real" data to a representation with enough precision to discriminate the relevant features?
- Filtering
  - Are we retaining/removing the "important/non-relevant" structures of the data?
  - Frequency/spatial domain filtering (noise, clipping/cropping)
- Selecting the "right" variable
  - Does this variable reflect the interesting features?
  - Does this variable allow for a "critical point" analysis?



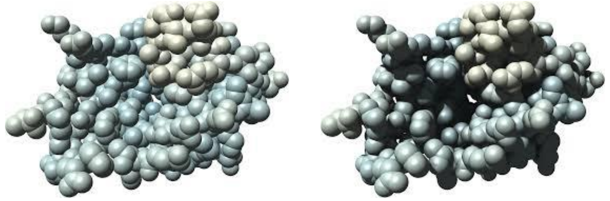


Sources of Error

- Functional model for resampling
  - What kind of information do we introduce by interpolation and approximation?
- Mapping
  - Are we choosing the graphical primitives appropriately in order to depict the kind of information we want to get out of the data?
  - Think of some real world analogue (metaphor)

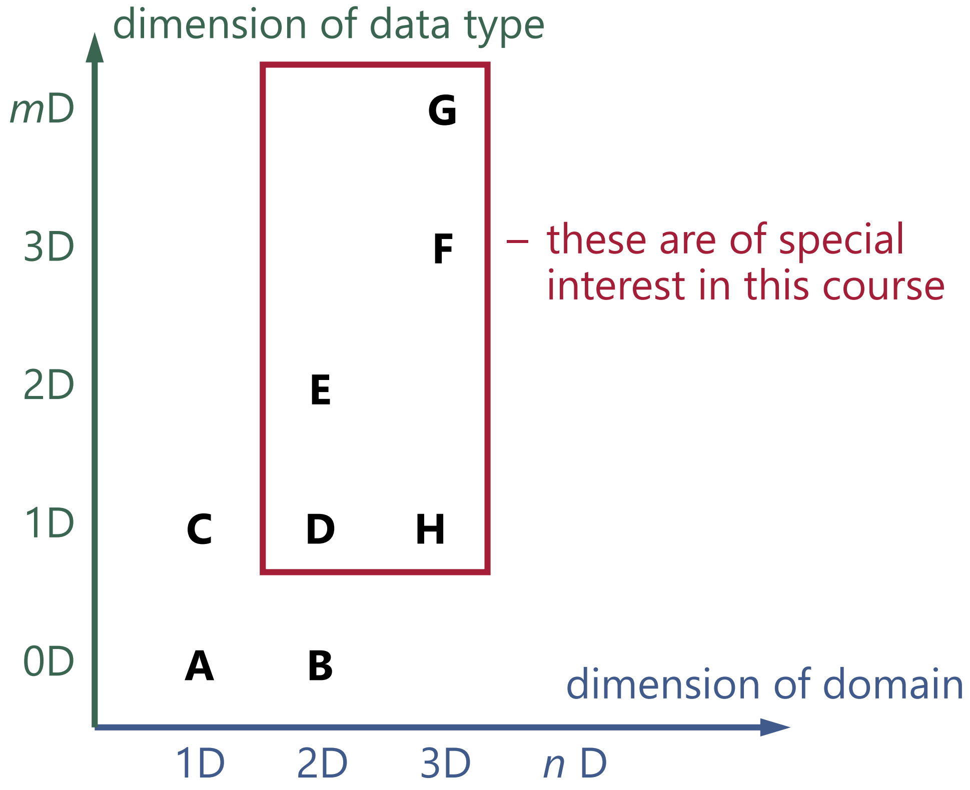


- Rendering
  - Need for interactive rendering often determines the chosen abstraction level
  - Consider limitations of the underlying display technology
  - Carefully add “realism”
    - The most realistic image is not always the most informative one!



Data Representation

- Classification of visualization techniques according to
  - Dimension of domain (independent variables)
  - Type and dimension of data (dependent variables)



Examples:

A: gas station along a road
B: map of cholera in London
C: temperature along a rod
D: height field of a continent
E: 2D air flow
F: 3D air flow in the atmosphere
G: stress tensor in a mechanical part
H: ozone concentration in the atmosphere

Data Representation

- Various application domains (e.g., engineering, natural sci., medicine…)
- Mostly measured or simulated data

→ **Numerical data sets**

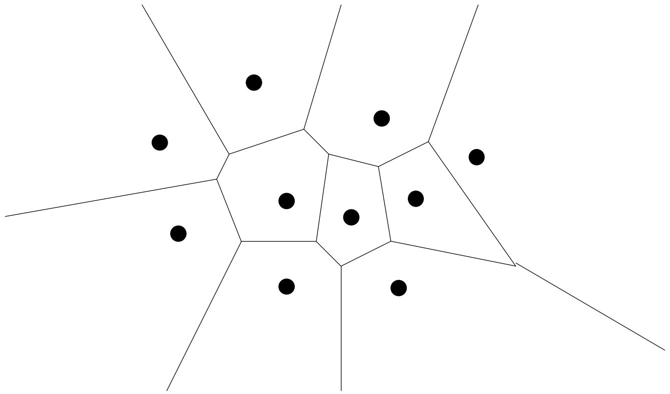
- Characteristics of data sets
  - Dimension of domain: number of coordinates or parameters
  - Dimension of values: scalar, vector, tensor, multivariate
  - Discretized data (grids represent continuous fields)
    - Type of discretization: (un-)structured grid, scattered data, …
  - Static vs. time-dependent (values and/or discretization)

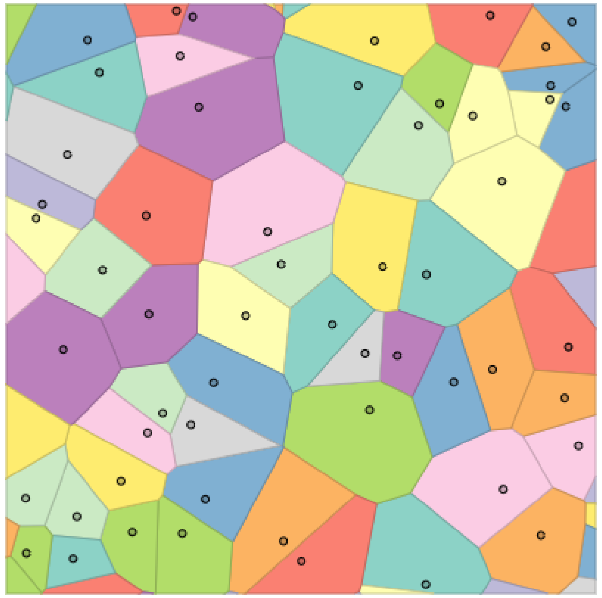
Domain

- Influence of data points
  - Only values at sample points within in a certain region influence an arbitrary point or all samples have to be considered
- Local influence
  - Only within a certain region
    - Voronoi diagram
    - Cell-wise interpolation (see later)
- Global influence
  - Each sample might influence any other point within the domain
    - Material properties for whole object
    - Scattered data interpolation

Domain

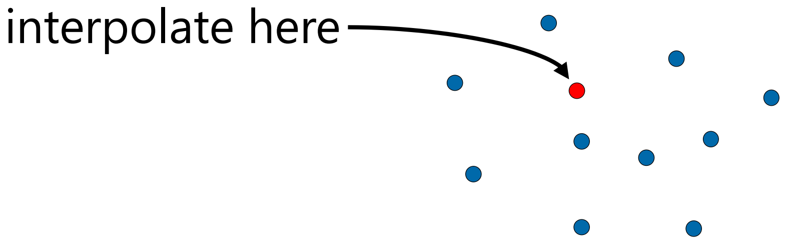
- Voronoi diagram
  - Construct a region around each sample point that covers all points that are closer to that sample than to every other sample
  - Each point within a certain region can get assigned the value of the sample point (nearest neighbor interpolation)

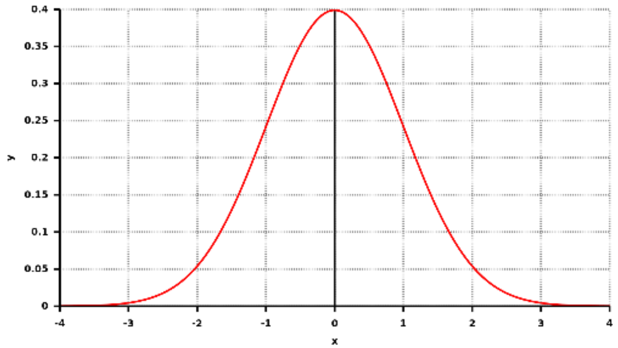




Domain

- Scattered data interpolation
  - At each point the weighted average of all sample points in the domain is computed
  - Weighting functions determine the support of each sample point
    - Radial basis functions simulate decreasing influence with increasing distance from samples (e.g., Gaussian distribution)
  - Schemes might be non-interpolating and expensive in terms of numerical operations





Gaussian distribution

Data Structures

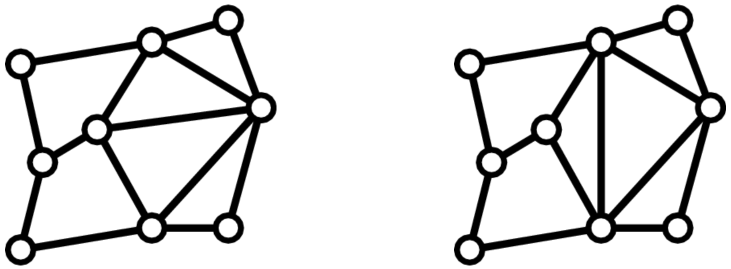
- Requirements:
  - Efficiency of accessing data
  - Space (memory) efficiency
  - Portability
    - e.g. *Binary* (less portable, more efficient) vs. *Text* (human readable, portable, less efficient)
- Definition
  - If points are arbitrarily distributed and no connectivity exists between them, the data is called *scattered*
  - Otherwise, the data is composed of cells with common boundaries
  - **Topology** specifies the structure (*connectivity*) of the data
  - **Geometry** specifies the shape (*position*) of the data

Data Structures

- Some definitions concerning topology and geometry
  - Topology is concerned with qualitative questions about geometrical structures
    - Does it have any holes in it?
    - Is it all connected together?
    - Can it be separated into parts?
  - Underground map does not tell you how far one station is from the other, but rather how the lines are connected (topological map)

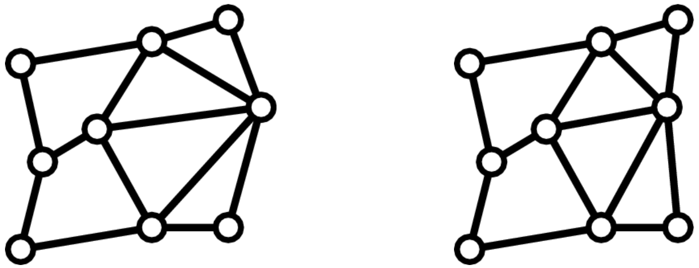
Data Structures

- Topology
  - Properties of geometric shapes that remain unchanged under smooth deformation



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

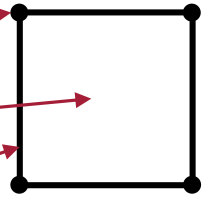
  - Topologically equivalent
    - Things that can be transformed into each other by stretching and squeezing, without tearing or sticking together bits previously separated



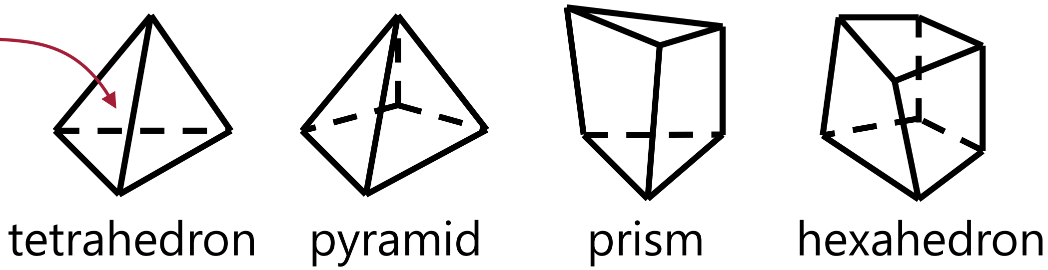
Topologically equivalent

Data Structures

- Grid elements
  - Nodes
  - Cells
  - Edges

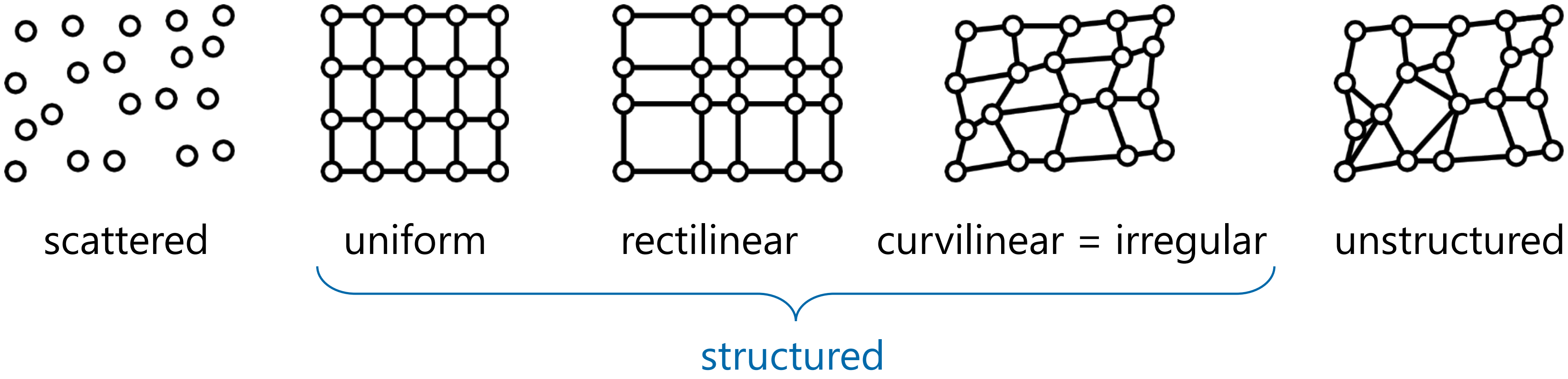


  - Faces (3D)



- Cell types
  - Tetrahedra, pyramids, prisms, hexahedra, …
  - Quadrilateral faces are non-planar in general!
- Grid data
  - Node-based (value per node): e.g. fluid dynamics simulations
  - Cell-based (value per cell): e.g. medical scanner data
  - Edge-/Face-based (value per edge/face, rare): e.g. higher-order elements

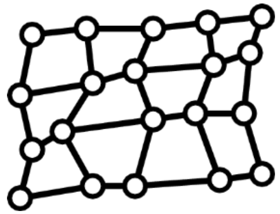
Data Structures – Common Grid Types



- Scattered data: no connectivity, often result from measurements
- Uniform, rectilinear, curvilinear grids are structured grids:
  - Structured grids have regular (implicit) topology & regular/irregular geometry
- Unstructured grids:
  - Have irregular topology & reg./irreg. geom., cell type may vary (tets, prisms,...)
  - Topology and cell type have to be stored explicitly

Data Structures

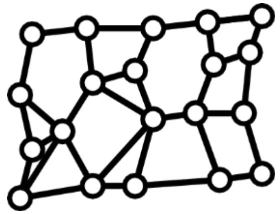
- Characteristics of **structured grids**
  - Easier to compute with
  - Often composed of sets of connected parallelograms (hexahedra), with cells being equal or distorted with respect to (non-linear) transformations
  - May require more cells or badly shaped cells in order to precisely cover the underlying domain
  - Topology is represented implicitly by an *n*-vector of dimensions
  - Geometry may be represented explicitly by an array of points
  - Every interior point has the same number of neighbors



structured

Data Structures

- If no implicit topological (connectivity) information is given, the grids are called **unstructured grids**
  - Sometimes obtained by triangulation of points sets
  - Often designed with respect to the used simulation solver
- Characteristics of **unstructured grids**
  - Grid point geometry and connectivity (and cell type) must be stored
  - Dedicated data structures needed to allow for efficient traversal and, thus, data retrieval
  - Often composed of tetrahedra or hexahedra
  - Typically, fewer cells are needed to cover the domain
  - Cells have to be convex for many visualization techniques!

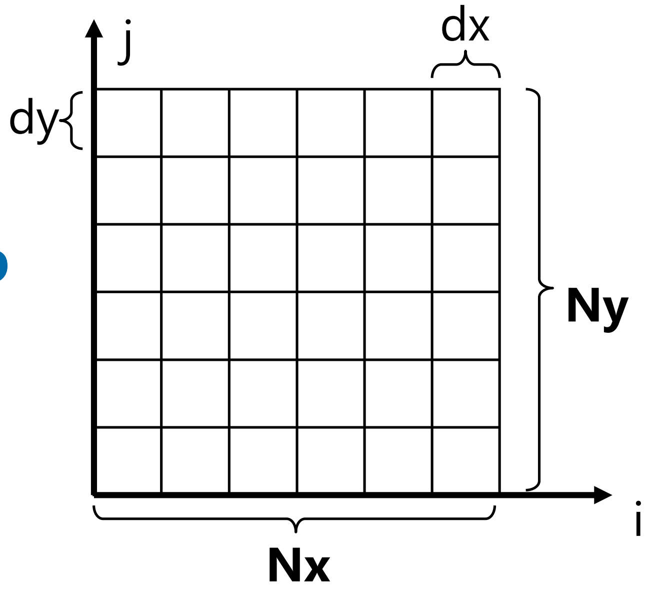


unstructured

Data Structures

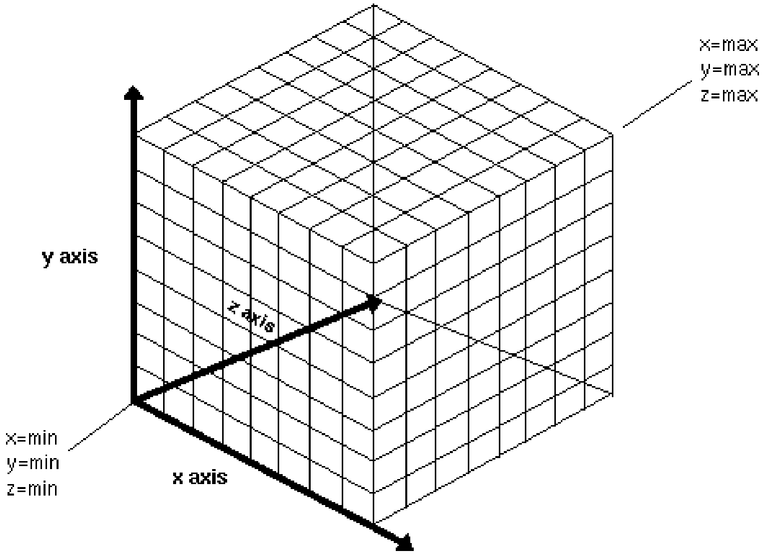
- Cartesian or equidistant grids
  - Structured grid
  - Cells and nodes are numbered sequentially with respect to increasing x, then y, then z, or vice versa
  - Number of nodes = Nx•Ny•Nz
  - Number of cells = (Nx - 1)•(Ny - 1)•(Nz - 1)
  - Cell size: dx = $(x_{max} - x_{min}) / (Nx-1)$, … (node-based data)

2D



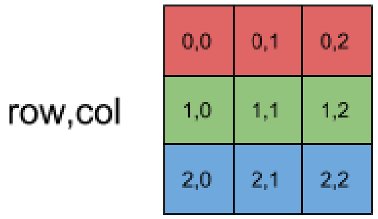
dx = dy = dz

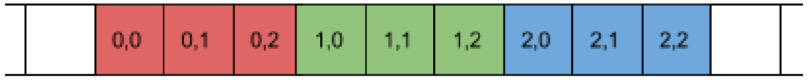
3D

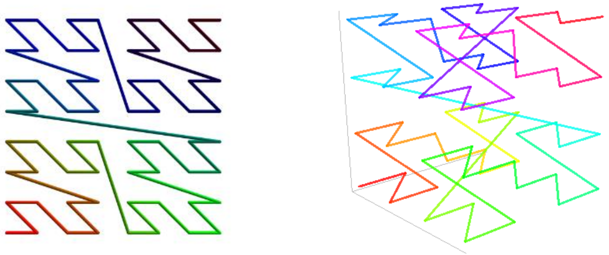


Data Structures

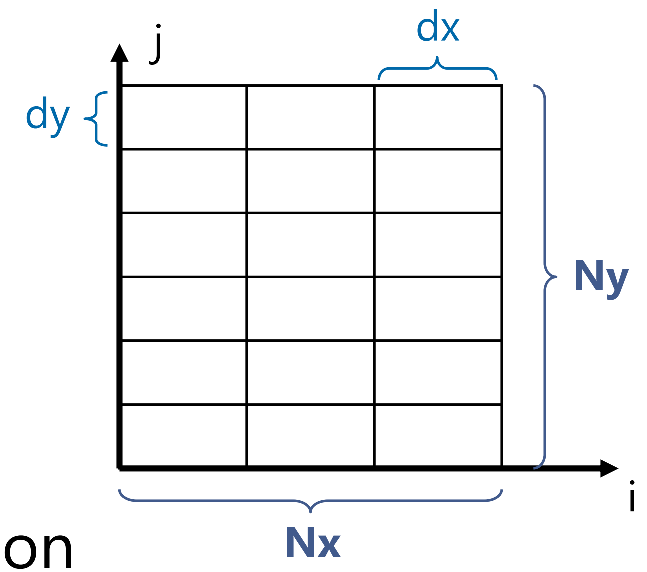
- Cartesian grids
  - Vertex positions are given implicitly from [i,j,k]:
    - $P[i,j,k].x = origin\_x + i \cdot dx$
    - $P[i,j,k].y = origin\_y + j \cdot dy$
    - $P[i,j,k].z = origin\_z + k \cdot dz$
  - Global vertex index $I[i,j,k] = k \cdot Ny \cdot Nx + j \cdot Nx + i$
    - $k = I / (Ny \cdot Nx)$
    - $j = (I \% (Ny \cdot Nx)) / Nx$
    - $i = (I \% (Ny \cdot Nx)) \% Nx$
  - Global index allows for linear storage scheme
    - Wrong access pattern might destroy cache coherence
    - Sometimes other mapping to linear memory layout for better cache coherence, e.g. space-filling curves







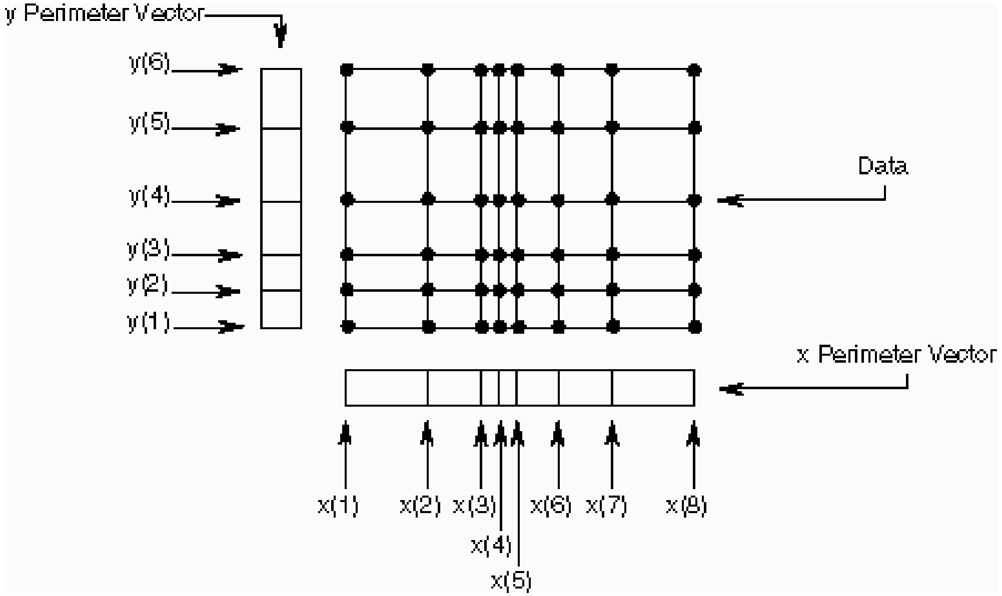
Data Structures



- Uniform grids
  - Similar to Cartesian grids
  - Consist of equal cells but with different resolution in at least one dimension (dx ≠ dy (≠ dz))
  - Spacing between grid nodes is constant in each dimension
    → same indexing scheme as for Cartesian grids
  - Most likely to occur in applications where the data is generated by a 3D imaging device providing different sampling rates in each dimension
  - **Typical example:**
    medical volume data consisting of slice images
    - Slice images with square pixels (dx = dy)
    - Larger slice distance (dz > dx = dy)

Data Structures

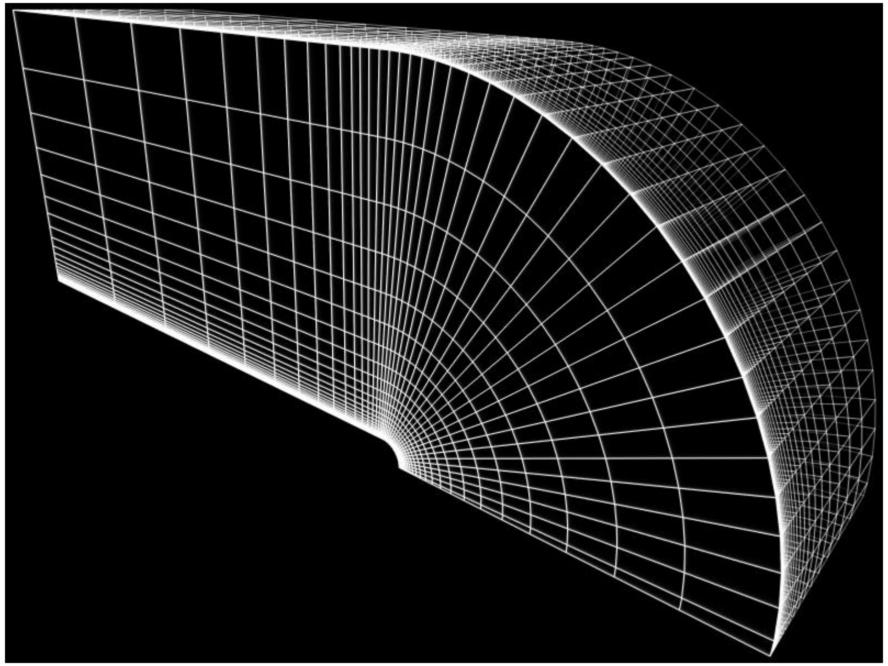
- Rectilinear grids
  - Topology is regular but with irregular spacing between grid nodes
    - Non-linear scaling of positions along either axis
    - Spacing, x_coord[L], y_coord[M], z_coord[N], must be stored explicitly
  - Topology still is implicit

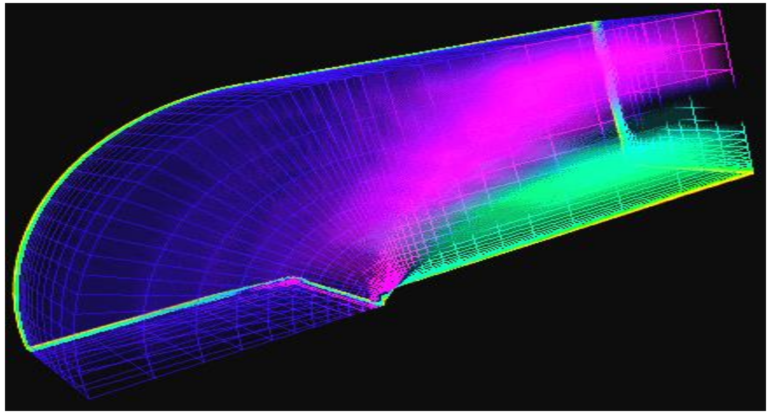


(2D perimeter lattice:
rectilinear grid in IRIS Explorer)

Data Structures

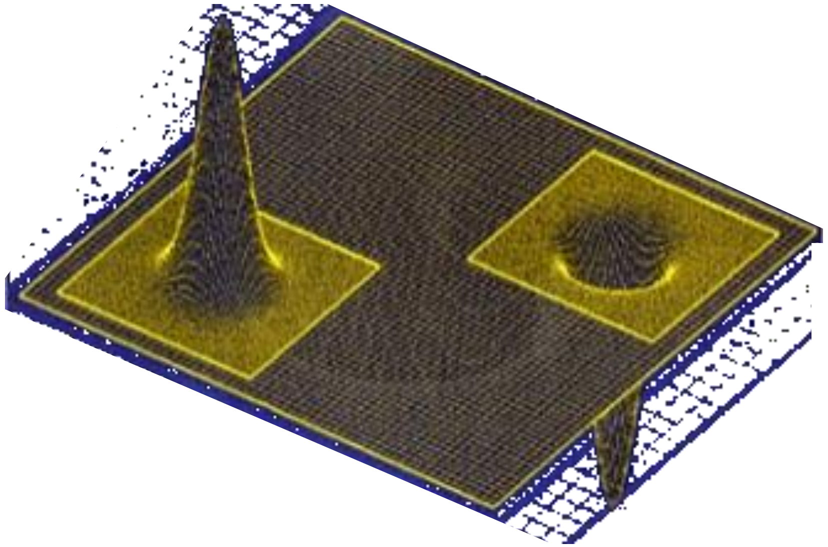
- Curvilinear (irregular) grids
  - Topology is regular but with irregular spacing between grid nodes
    - Positions are non-linearly transformed
  - Topology still is implicit, but node positions are explicitly stored
    - x_coord[L,M,N]
    - y_coord[L,M,N]
    - z_coord[L,M,N]
  - Geometric structure might result in concave grids
    - Difficulties, e.g.; with Sorting, ray intersection, …
  - Cell edges are straight, not curved

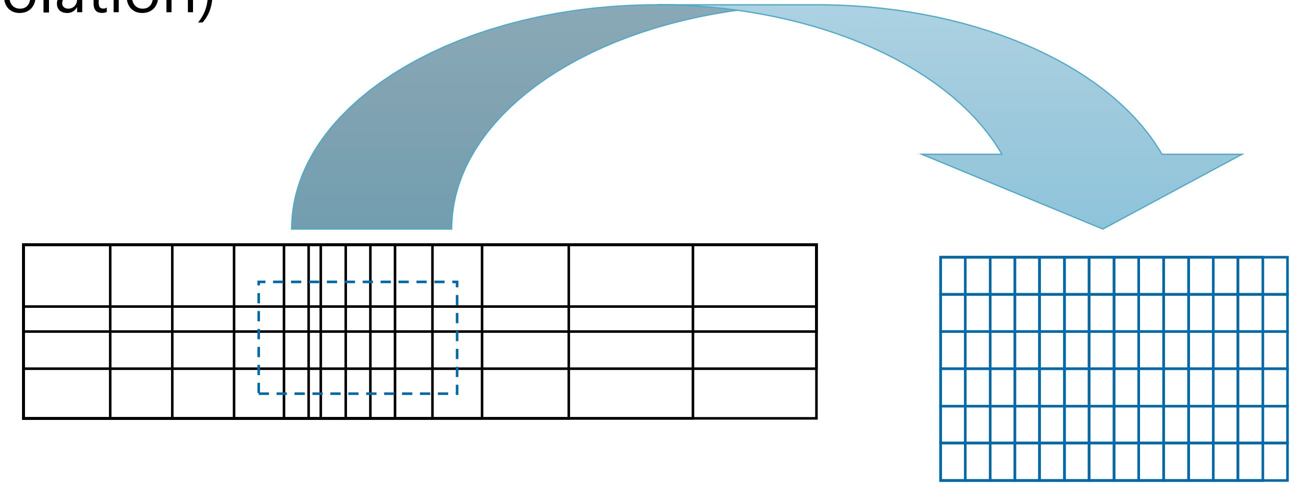




Data Structures

- Multigrids
  - Focus in specific areas to avoid unnecessary detail in other areas
  - Finer grid for regions of interest
  - Difficulties in the boundary region (i.e., interpolation)



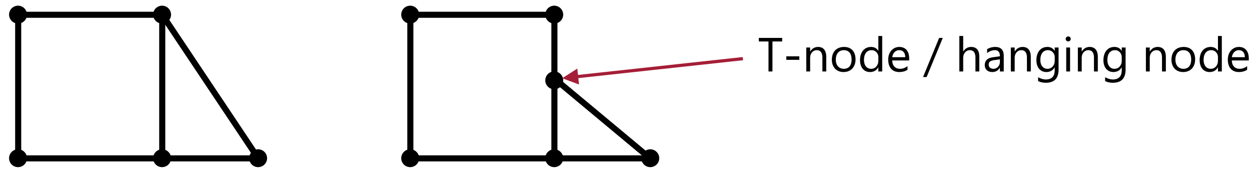


Data Structures

- Characteristics of structured grids
  - Structured grids can be stored in a 2D / 3D array
  - Arbitrary samples can be directly accessed by indexing
  - Topological information is implicitly coded
    - Direct access to neighbor cells
  - Cartesian, uniform, and rectilinear grids are necessarily convex
    - Rigid layout prohibits geometric structure to adapt to local features
  - Curvilinear grids are a more flexible alternative to model arbitrarily shaped objects, but might be concave
    - Sorting of grid elements is a more complex procedure
    - Cells have to be convex for most visualization techniques!

Data Structures

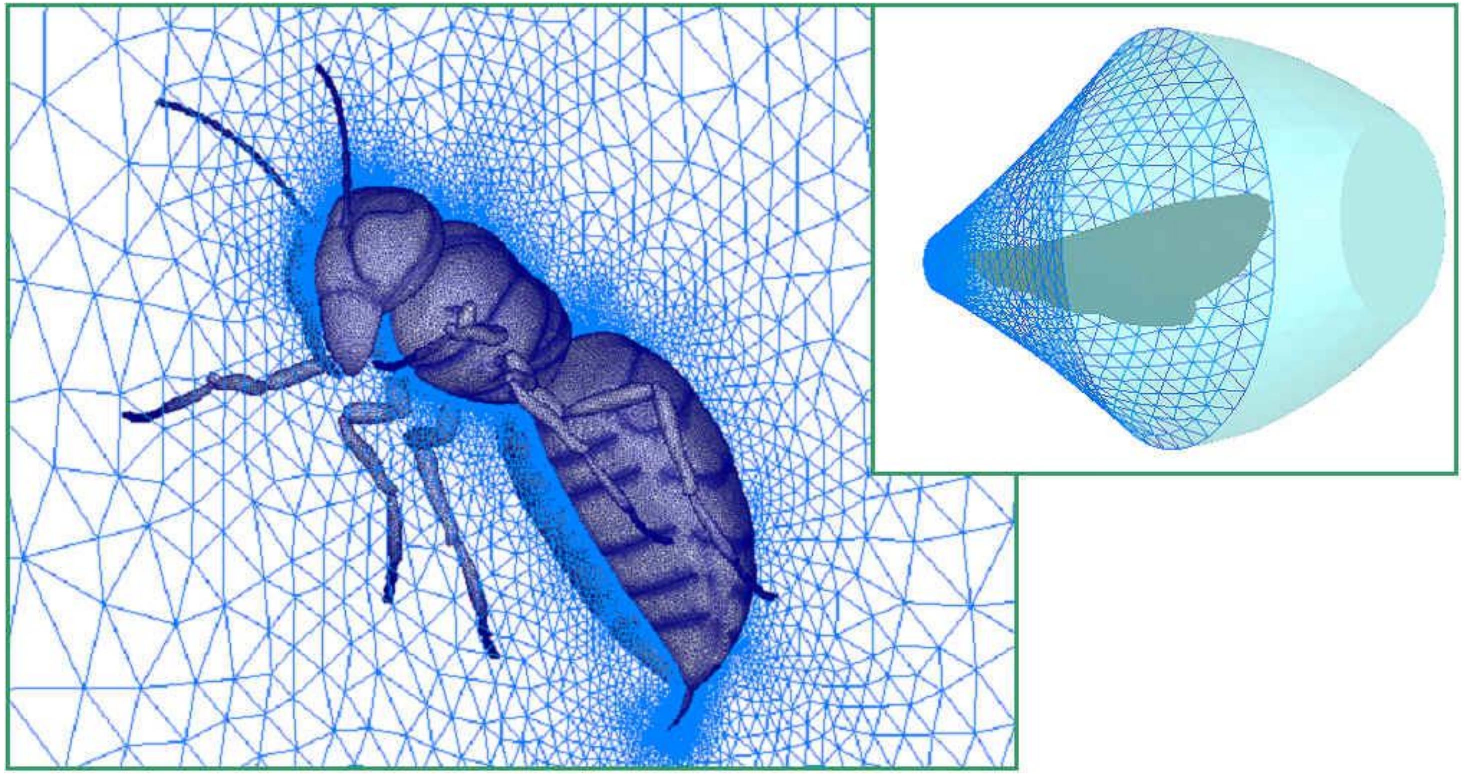
- Unstructured grids
  - Composed of arbitrarily positioned and connected cells
  - Can be composed of one unique cell type or can be hybrid/mixed
    - Tetrahedra, pyramids, prisms, hexahedra,…
  - Cells can include 1D (lines) or 2D (triangles, quads) also in 3D domains, often used for defining boundary conditions in simulations
  - Adjacent cells have to share faces, edges and nodes
    - No T-nodes / hanging nodes

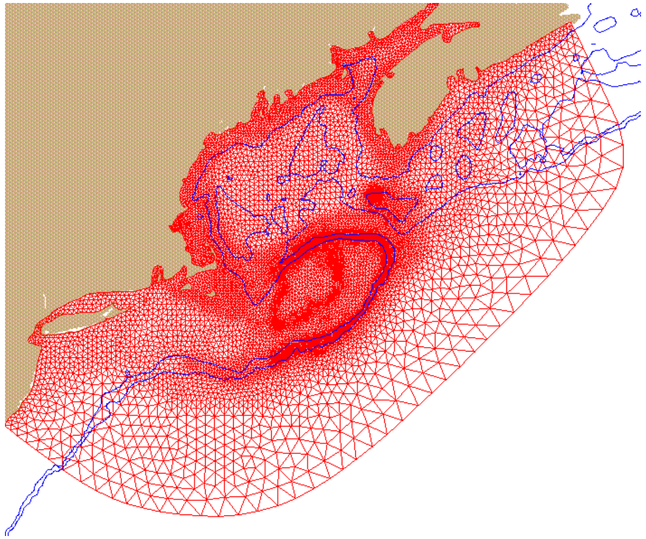


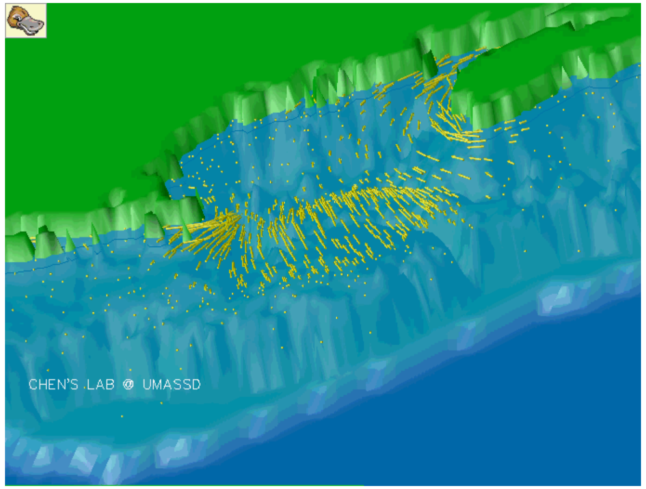
  - Can adapt to local features (small vs. large cells) → adaptive refinement

Data Structures – Unstructured Grids

- **Examples:** Adaption to local features

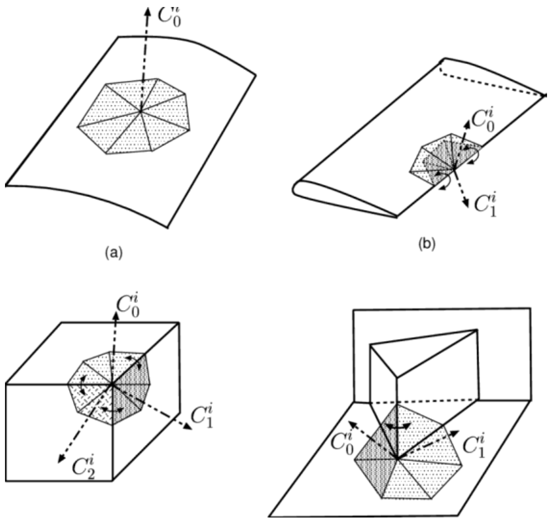




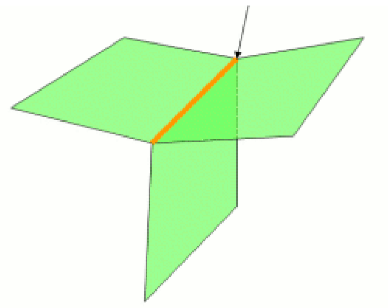


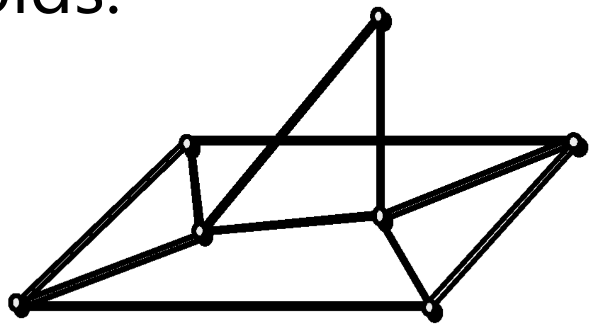
Data Structures

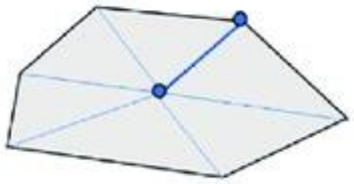
- Manifold meshes
  - 2-manifold is a surface where at every point on the surface a surrounding area can be found that looks like a disc
  - Everything can be flattened out to a plane
  - Sharp creases and edges are possible needs more than one normal per vertex
    (a distinguished normal for each cell at the vertex)



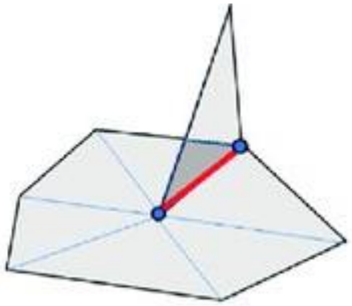
- Examples of non-manifolds:







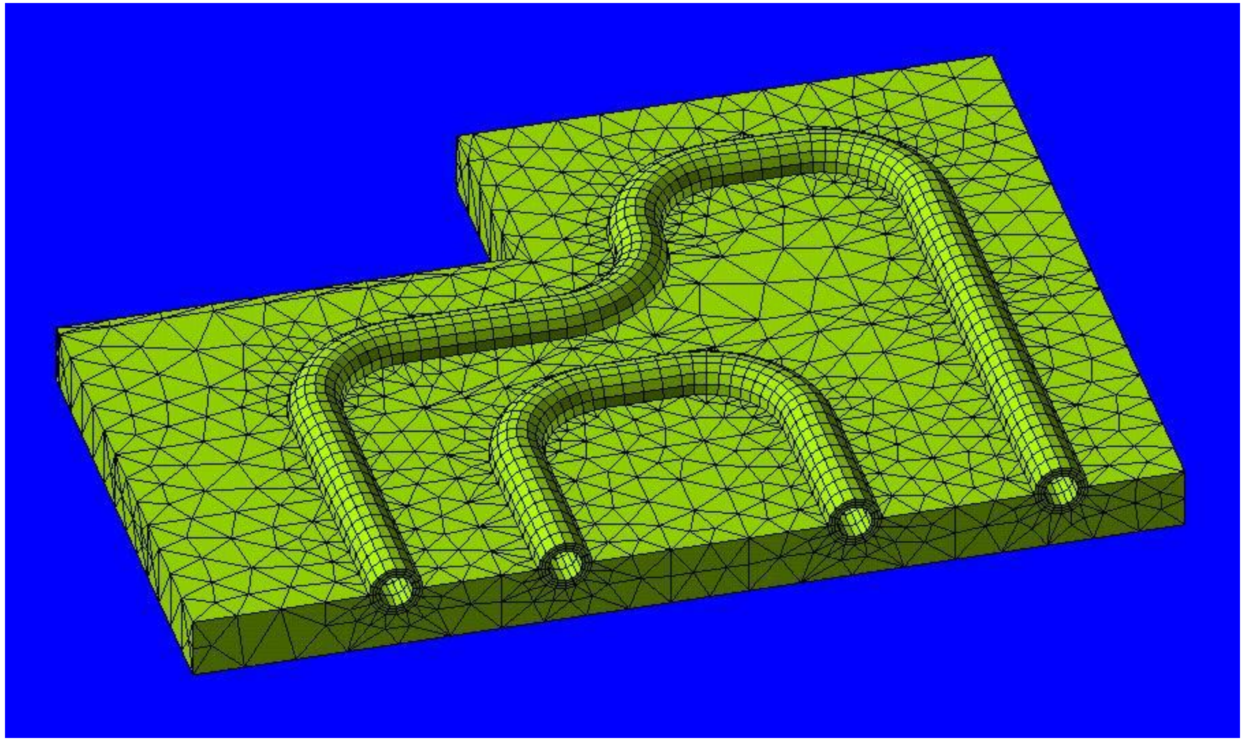
(b) manifold edge



(c) non-manifold edge

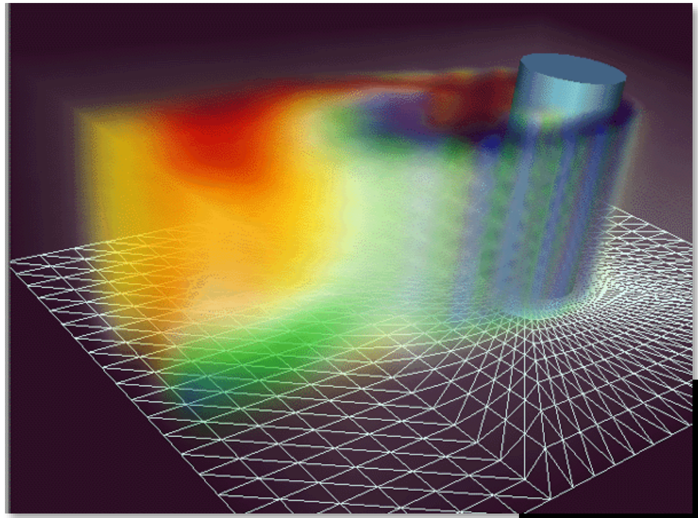
Data Structures

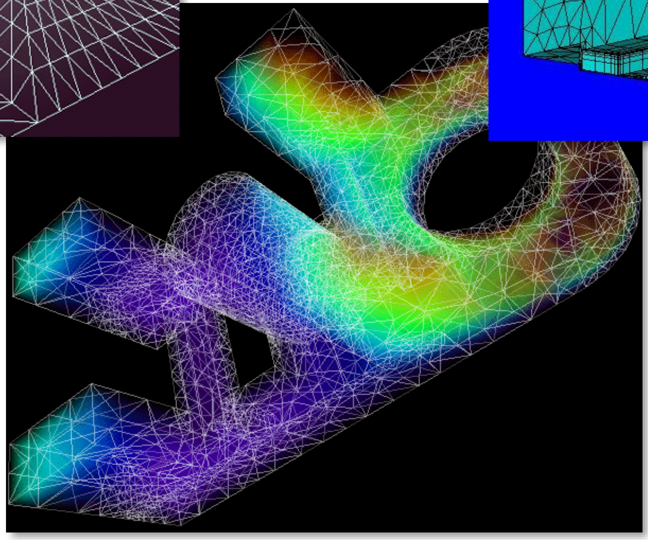
- Hybrid grids
  - Combination of different grid types
  - Here: one tetrahedral unstructured grid and two hexahedral structured grids

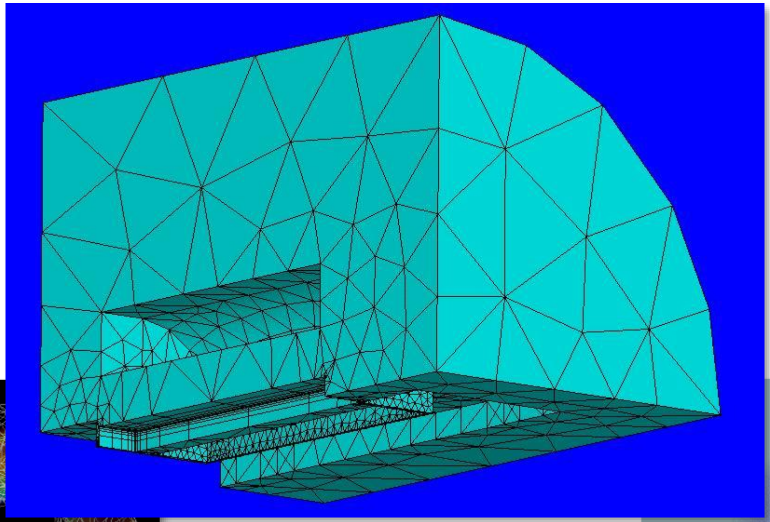


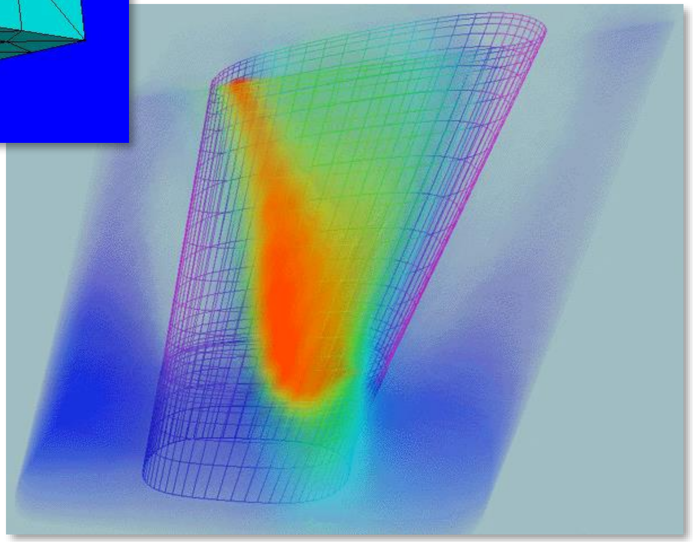
Data Structures

- Examples



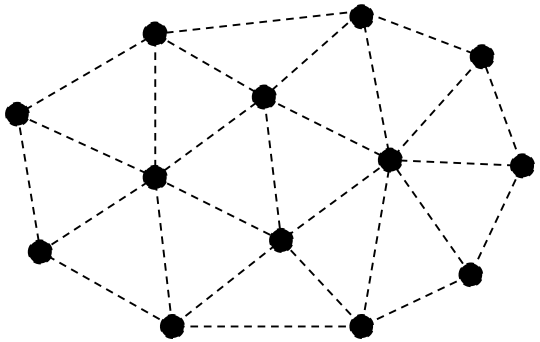






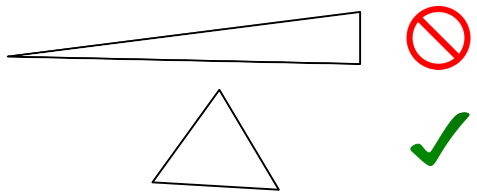
Data Structures

- Scattered data
  - Irregularly distributed positions without connectivity information
  - To get connectivity find a “good” triangulation
    (triangular/tetrahedral mesh with scattered points as nodes)

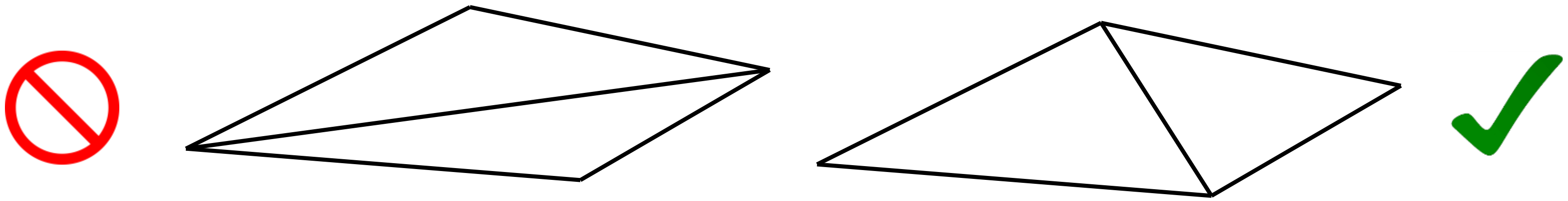


Data Structures

- For a set of points there are many possible triangulations
  - A measure for the quality of a triangulation is the aspect ratio of the triangles
  - Avoid long, thin ones
  - Delaunay triangulation (→ later in the course)



$$\frac{\textit{radius of incircle}}{\textit{radius of circumcircle}} \quad \textit{or} \quad \frac{\textit{maximum}}{\textit{minimum}}\ \textit{angle in triangle}$$



Data Values

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

Data Values

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

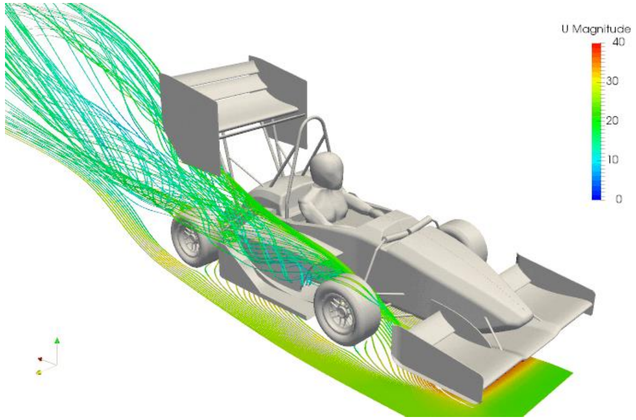
Data Values

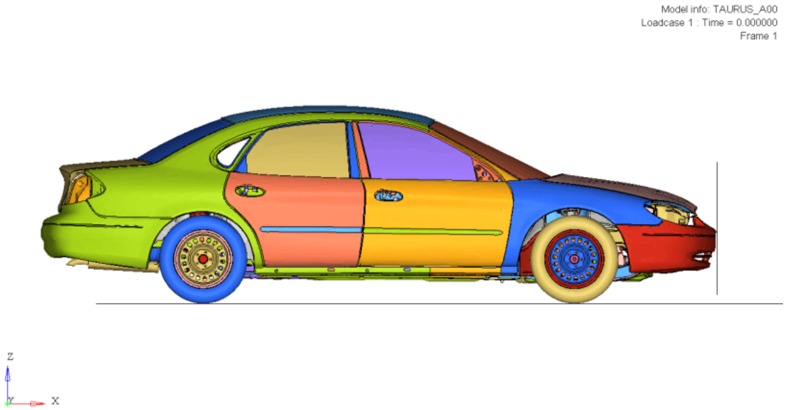
- Data types
  - **Scalar data** is given by a function $f(x_1, \ldots, x_n): \mathbb{R}^n \rightarrow \mathbb{R}$ with n independent variables x_i
  - **Vector data**, representing direction and magnitude, is given by an m-tuple $(f_1, \ldots, f_m)$ with $f_k = f_k(x_1, \ldots, x_n)$, $m \geq 2$ and $1 \leq k \leq m$
    - Usually $\mathrm{m} = n$
    - Exceptions, e.g., due to projection
  - **Tensor data**: a tensor of order k is given by $t_{i1,i2,\ldots,ik}(x_1, \ldots, x_n)$ (→ later in the course)
    - A tensor of order 0 is a scalar, order 1 is a vector, of order 2 is a matrix, …

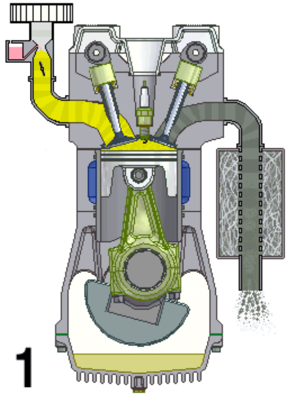
- Structure of the data
  - Sequential (in the form of a list)
  - Relational (as table)
  - Hierarchical (tree structure)
  - Network structure

Data Classification

- Number/type of independent and dependent variables
- Time dependency
  - 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 of an airplane
    - Data and grid geometry (topology remains constant), e.g., crashworthiness of cars
    - Data, grid geometry and topology, e.g., engine simulation with moving piston

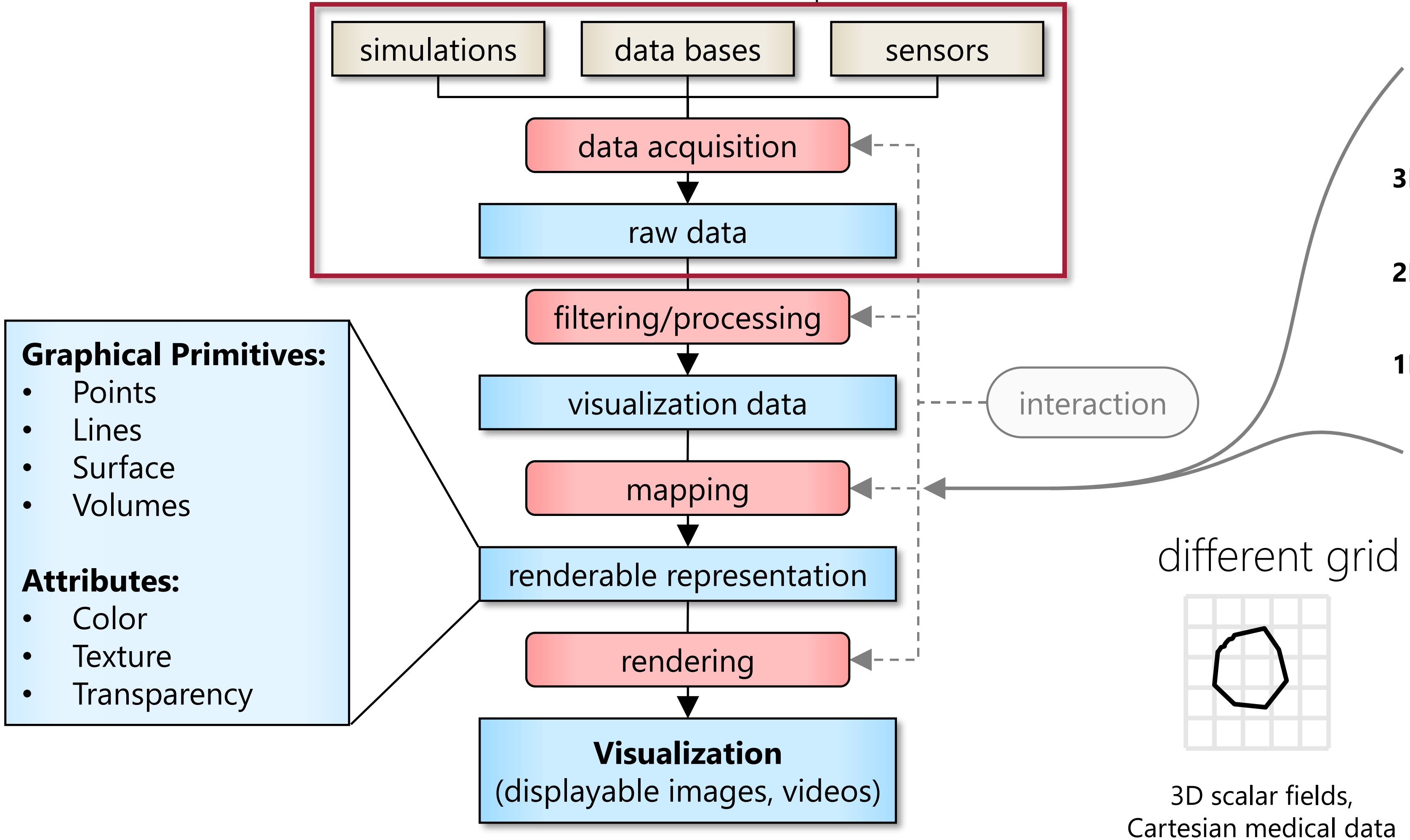




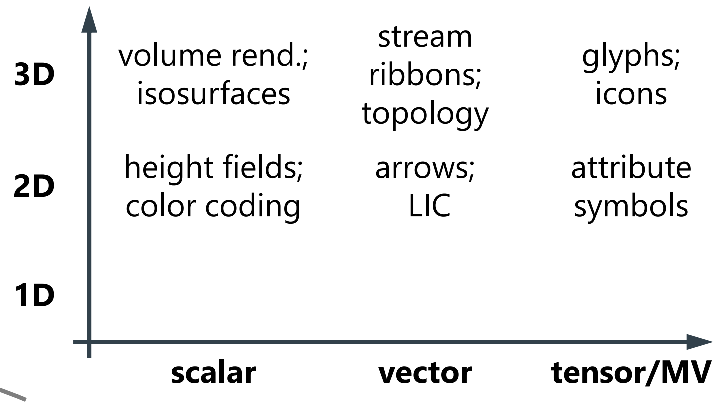


Classification of Visualization Methods

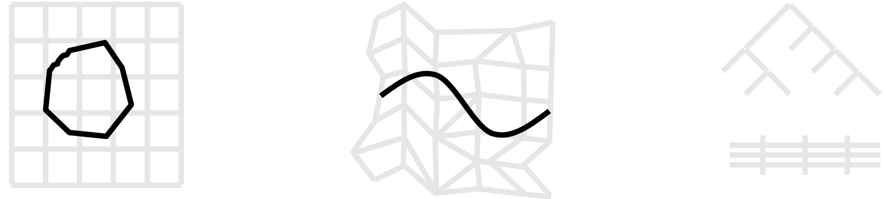
The Visualization Pipeline



Mapping – classification



different grid types → different algorithms



3D scalar fields, Cartesian medical data

3D vector fields un/structured, CFD

trees, graphs, tables, data bases, InfoVis