Part II All About Data

Data Abstraction

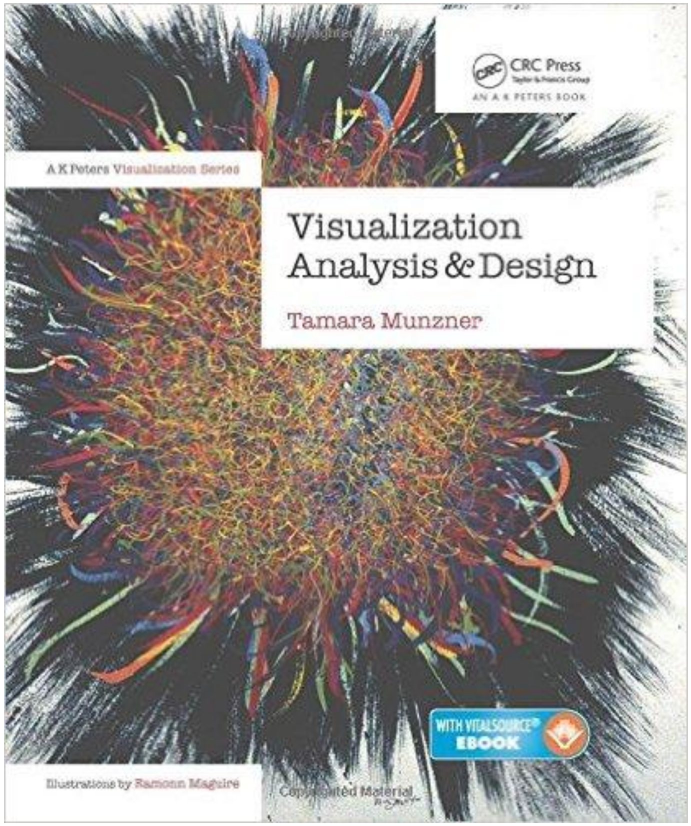
The scientist described what is: the engineer creates what never was.

Theodor von Karman
The father of supersonic flight

Fall 2020

Source of This Unit

- **Material of this unit is based on Tamara Munzner, *Visualization Analysis and Design*, AK Peters/CRC Press, 2014.**



Overview

- ***Topics to be covered in this unit*:**
  - ***Types***
    - ***Data Types*: Item, Attribute, Link, Position, Grid, etc.,**
    - ***Dataset Types*: Table, Network, Field, Geometry, etc.**
    - ***Attribute Types*: Categorical, Ordered**
  - ***Data Semantics*: Key vs. Value, Temporal, etc.**

Data and Dataset: 1/5

- **The *type* of data is its structural or mathematical interpretation.**
  - **At the *data* level, it can be an item, a link, an attribute, etc.**
  - **At the *dataset* level, it is how these data types are combined into a larger structure such as a table, a tree, a field of values.**

Data and Dataset: 2/5

- **There are five basic data types:**
  - ***Item*: An individual entity that is discrete (e.g., a number, a row of a table, etc.)**
  - ***Attribute*: Some measurable property (e.g., salary, price, temperature, etc.)**
  - ***Link*: A relationship between items**
  - ***Position*: A spatial data (e.g., location, coordinates, etc.)**
  - ***Grid*: The strategy for sampling continuous data (e.g., geometric and/or topological) between cells.**

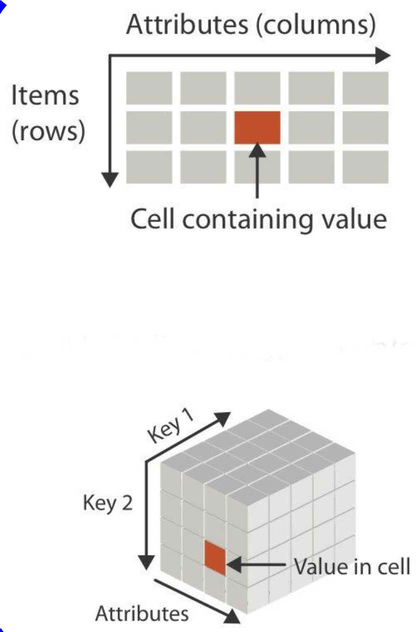
Data and Dataset: 3/5

- A ***dataset*** **is a collection of information that is the target of analysis.**
- **There are four basic types:**
  - **Tables**
  - **Networks**
  - **Fields**
  - **Geometry**
- **Complex combinations of multiple data types are commonly seen in real world applications.**

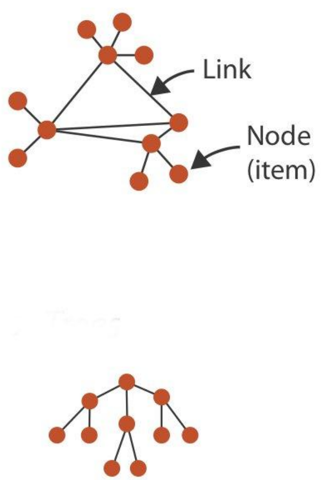
Data and Dataset: 4/5

- **There are four basic dataset types: tables, networks, fields and geometry.**

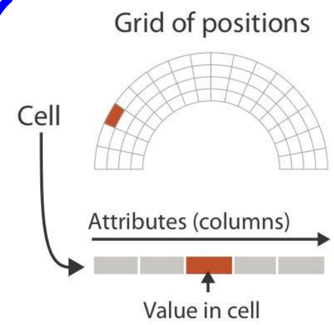
tables



networks



fields



geometry

Tamara Munzner, *Visualization Analysis and Design*, AK Peters, 2015

Data and Dataset: 5/5

- There are four basic dataset types: tables, networks, fields and geometry.
- There are other possible collections of items such as clusters, sets, lists, etc.

Tables	Networks & Trees	Fields	Geometry	Clusters, Sets, Lists
Items	Items (Node)	Grids	Items	Items
Attributes	Links	Positions	Positions	
	Attributes	Attributes		

Dataset Type (Table): 1/2

- **Tables are commonly seen type of datasets.**
- **In a 2D table, each row is an *item*, each column is an *attribute*, and each *cell* has a *value* of a particular item and a particular attribute.**

ID	Name	Addr.	City	State	Zip	Income	Phone
			....				
31765	John Dow	xxxxx	Houghton	MI	49931	50K	906-123-4567
			....				

item

attribute

cell

Dataset Type (Table): 2/2

- **A multidimensional table uses more indices.**
- **The following 3D table uses three indices to find a cell (i.e., Department, Gender, Status).**

Department		A	B	C	D	E	F
Male	Accepted	512	353	120	138	53	22
	Rejected	313	207	205	279	138	351
Female	Accepted	89	17	202	131	94	24
	Rejected	19	8	391	244	299	317

Dataset Type (Networks): 1/1

- **Networks (or graphs) are useful to represent relationship between several items.**
- **Here, an *item* is a *node* and a *link* is a relation between two items.**
- **Each node can have associated attributes (e.g., city size), and each link may also have associated attributes (e.g., distance between two cities).**
- **A tree is just a special type of networks.**

Dataset Type (Fields): 1/8

- **A *field* contains attribute values associated with cells.**
- **Each *cell* contains measurements or calculations from a *continuous* domain.**
- **Obtaining values from a continuous domain is usually very challenging because the domain is a continuum.**
- **A good *sampling* strategy for taking measurements from discrete positions is needed.**

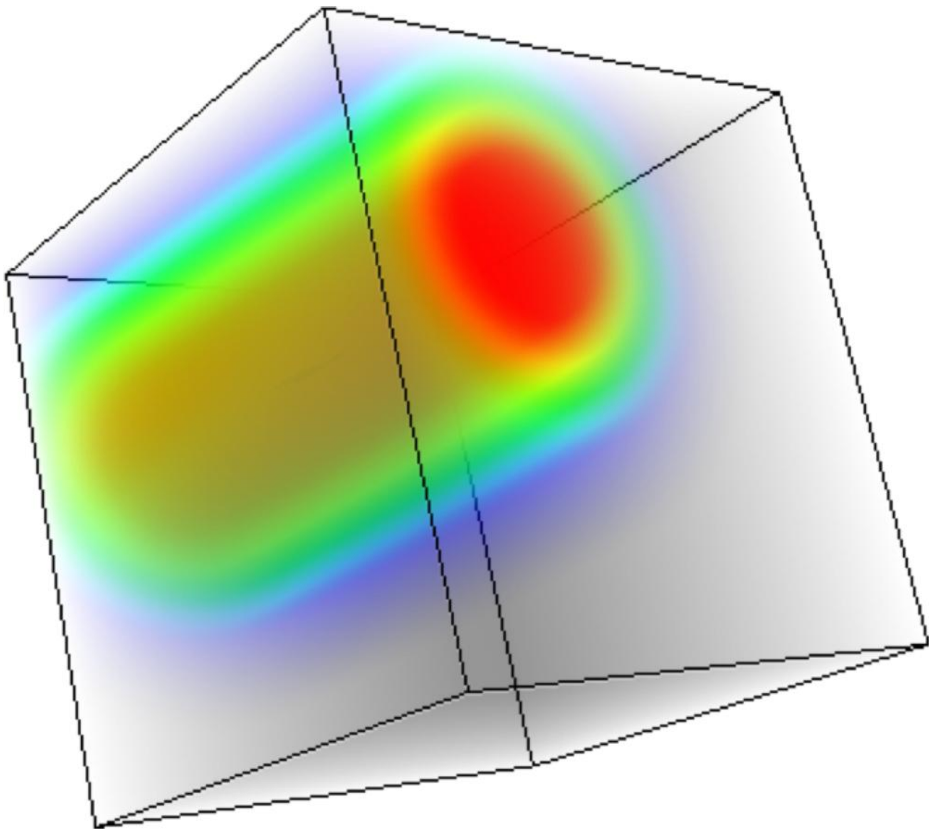
Dataset Type (Fields): 2/8

- **Because the number of measurements is only finite, we need to *interpolate* those missing measurements (i.e., showing the values between sampled values).**
- **With a good sample and an appropriate interpolation, the original continuum can be *reconstructed* so that the view is faithful to the measured values from an arbitrary viewpoint.**
- **The handling of continuous domain is always challenging.**

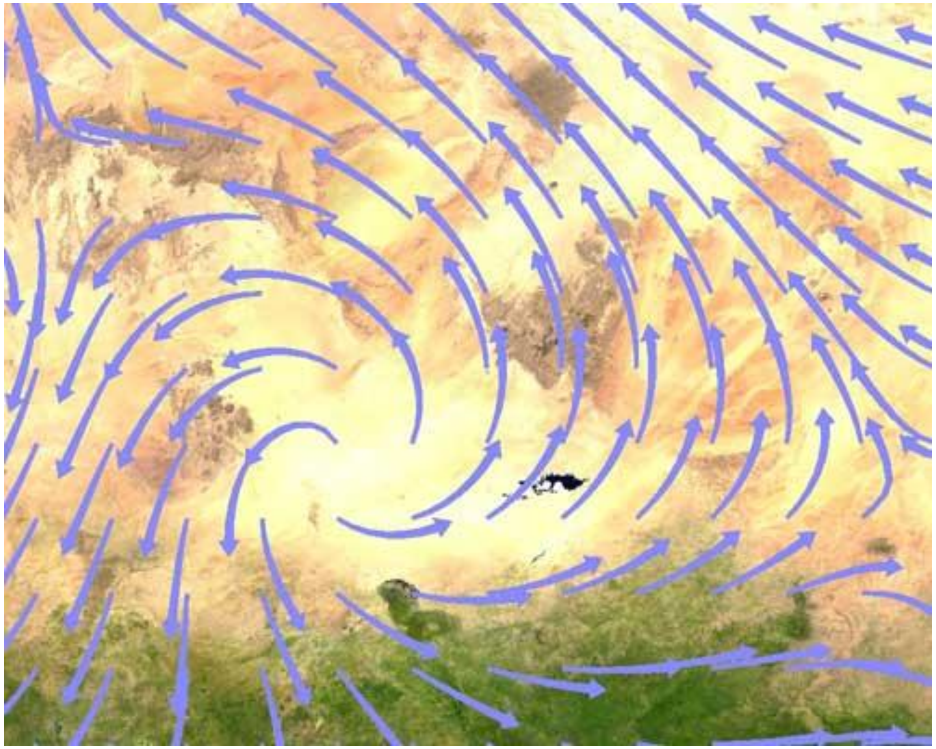
Dataset Type (Fields): 3/8

- **The cell structure of a *spatial field* is based on sampling at spatial positions.**
- ***Example*: We may measure the temperature at a space point and the result is represented by (x,y,z) – a point in space – and the measured temperature t. This is a *scalar field*, because the measurement is a single value.**
- ***Example*: We may also measure the velocity of a flow, and the result is a point (x,y,z) and a vector (i.e., velocity). This is a *vector field*.**

Dataset Type (Fields): 4/8



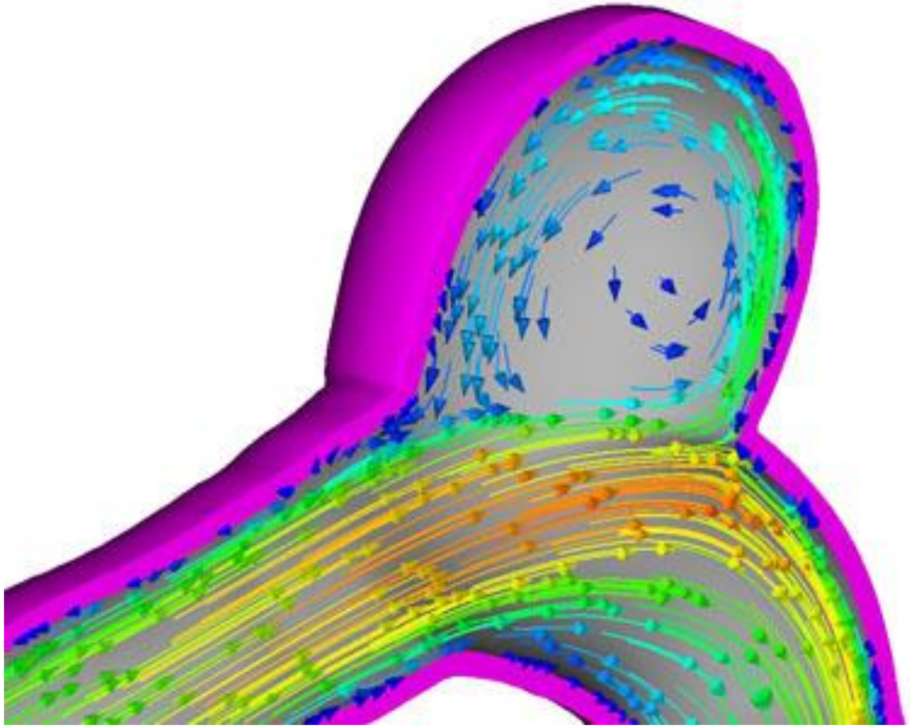
scalar field



vector field

Dataset Type (Fields): 5/8

- **The collected dataset may contain spatial information (i.e., the positions). This is a dataset that contains *spatial data*.**



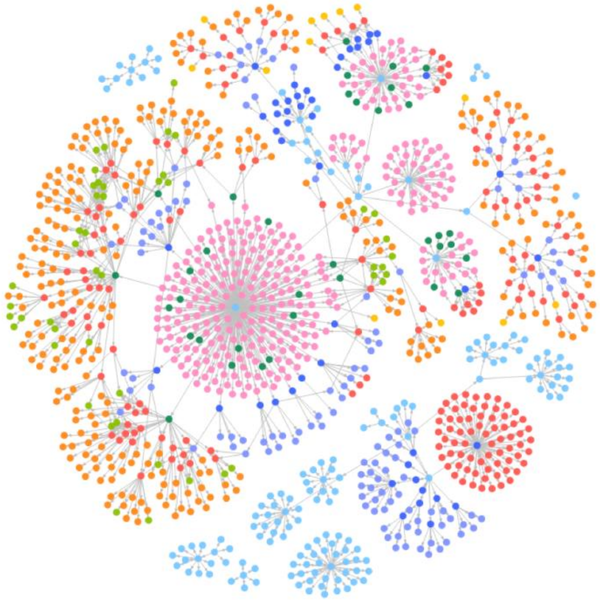
T.E. Tezduyar, S. Sathe, M. Schwaab and B.S. Conklin

Dataset Type (Fields): 6/8

- **In many applications, the spatial information may not be given. This dataset contains *non-spatial data*.**
- **Non-spatial data is sometimes also referred to as *abstract data*.**
- **In this case, the use of space is chosen by visualization designers.**

Dataset Type (Fields): 7/8

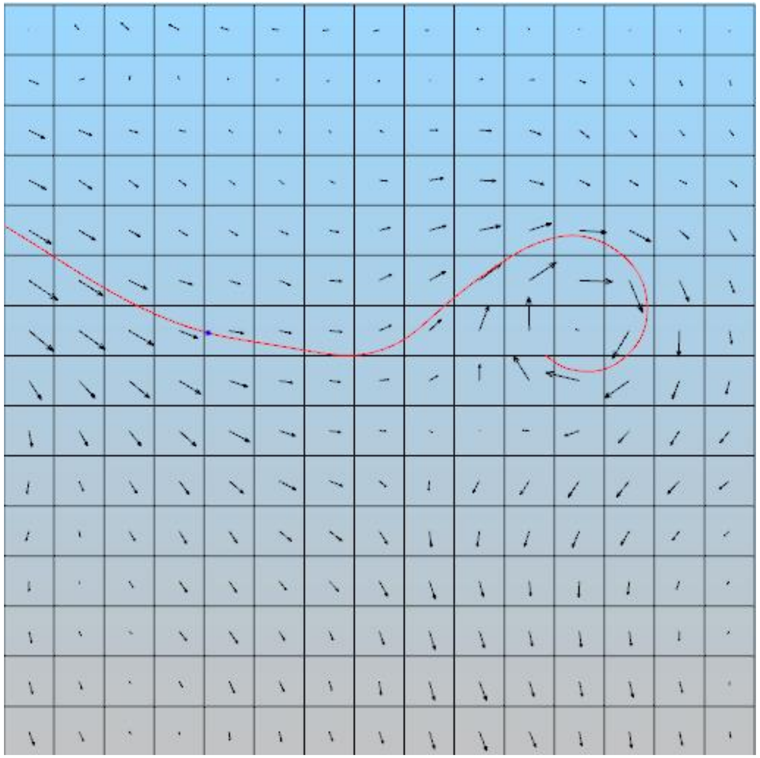
- ***Scientific visualization*** **is concerned with situation where spatial position is given, while in *information visualization* is concerned with situation where the use of space in a visual encoding is chosen by the designer.**



only the nodes and links are provided

Dataset Type (Fields): 8/8

- **When a field contains data created by sampling at a completely regular intervals, the cells form a *uniform grid*.**

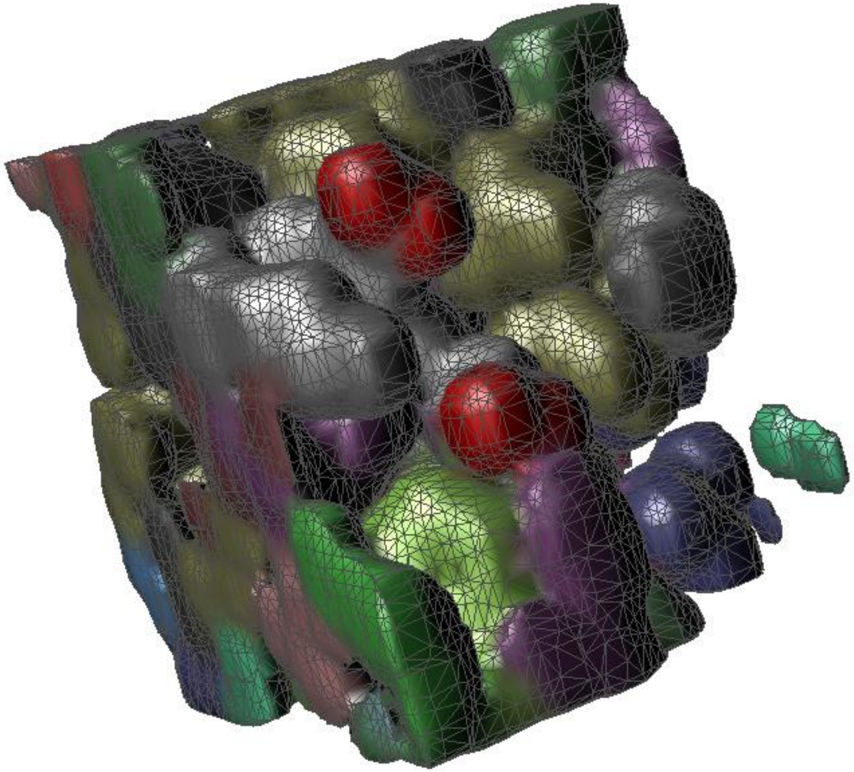


Geometry: 1/4

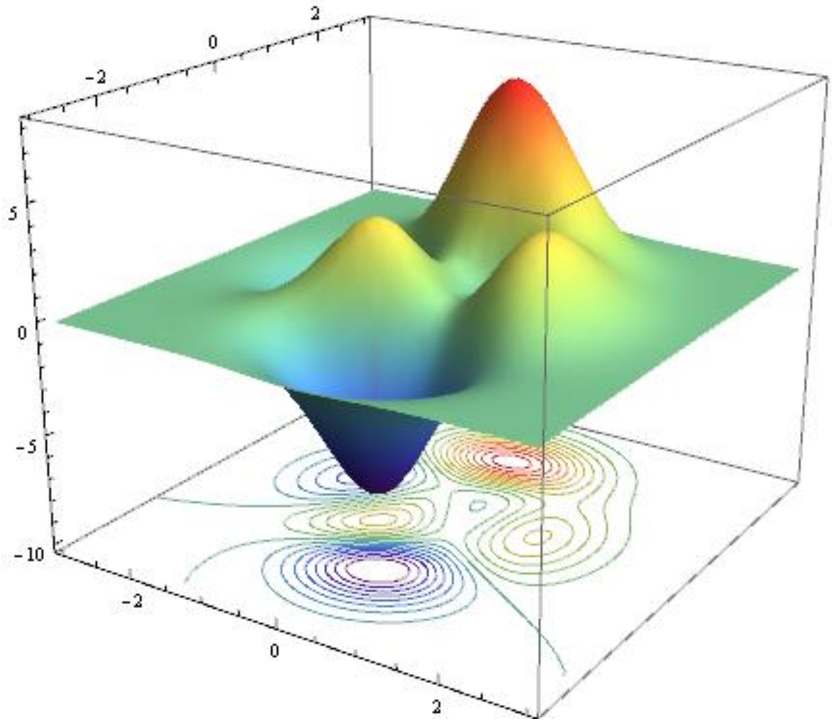
- The ***geometry*** dataset type specifies information about the shape of items with explicit spatial information.
- The items could be points, lines/curves, 2D surfaces/regions, 3D volumes, or even higher dimensional data.
- Therefore, geometry datasets are spatial, and typically occur in the context of tasks that require shape understanding.

Geometry: 2/4

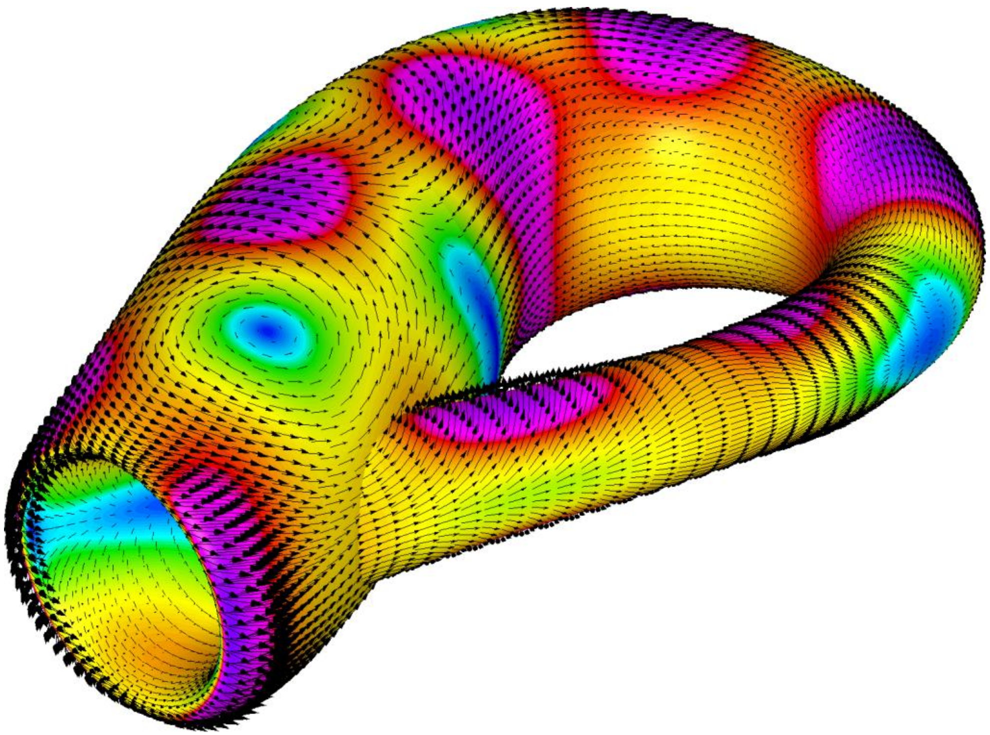
- **Geometry datasets may not have attributes. In situations where we only care about shape understanding, only the positions would be enough.**



Geometry: 3/4

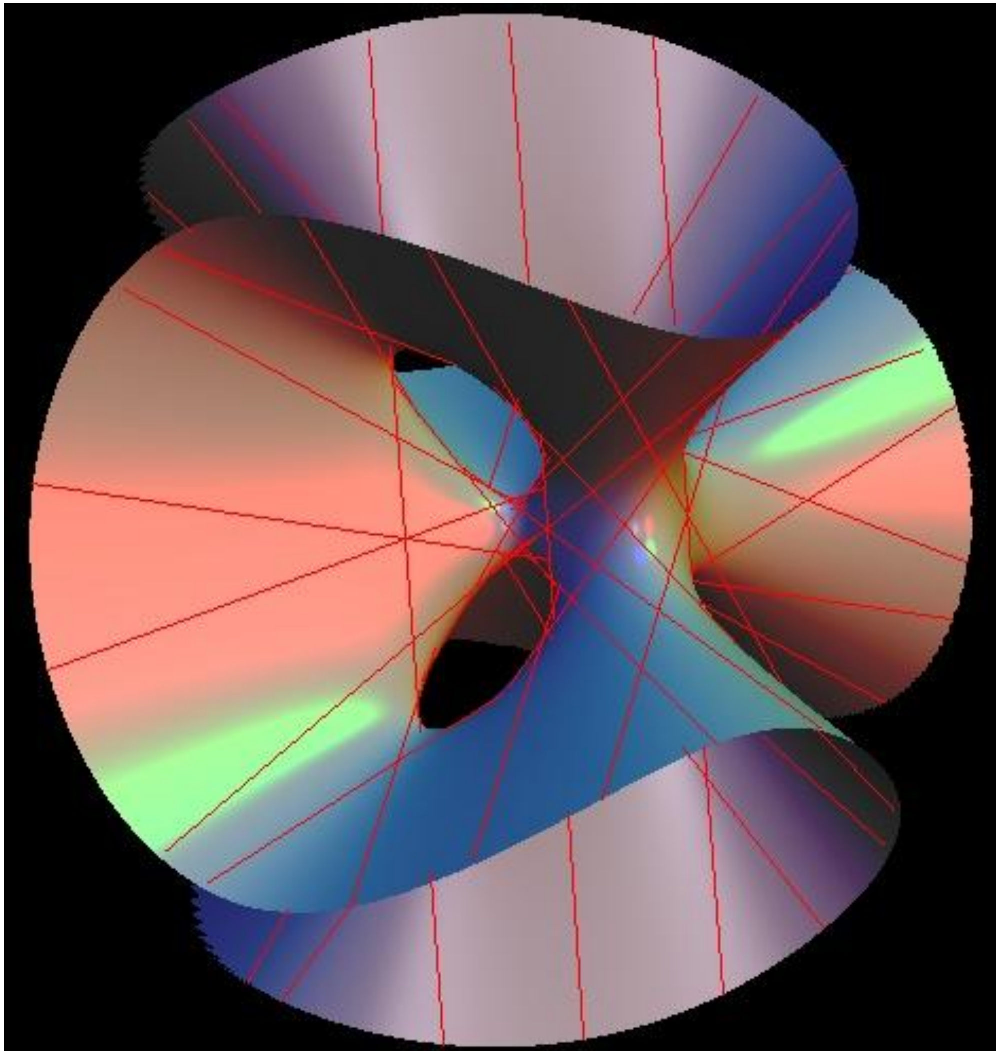


contours generated from
a spatial field



Klein bottle with a vector field
on a regular grid

Geometry: 4/4



Clebsch cubic surface with 27 lines

Dataset Availability

- **There are two kinds of dataset availability: *static* or *dynamic*.**
- **A *static file* means the entire dataset is available all at once.**
- **Some datasets are only available as *dynamic streams*, where the dataset information trickles in over the course of the visualization session.**
- **This dynamic information may mean adding new items, deleting existing ones, or the values of existing items may even change.**

Attribute Types

- **Attribute types are:**
  - **Categorical**
  - **Ordered**
    - **Ordinal**
    - **Quantitative**
  - **The direction of attribute ordering can be:**
    - **Sequential**
    - **Diverging**
    - **Cyclic**

Categorical Type

- ***Categorical*** **data does not have an implicit ordering, but often has a hierarchical structure.**
- ***Examples*****: Gender types, file types, shapes type (e.g., triangles, circles, rectangles, etc.), fruit types (e.g., apples, oranges, bananas, etc.)**
- **The above examples do not have an "implicit" order imposed to the data.**
- **However, one may enforce an order to each of the above example. For example, fruit names are arranged in alphabetical order.**

Ordered Type

- ***Ordered*** data has an implicit ordering. There are two ordered types: ordinal and quantitative.
  - ***Ordinal***: It has a well-defined ordering but cannot do full-fledged arithmetic.
    - ***Example***: shirt size, shoe size, grade (e.g., A, AB, etc.), ranking, zip code, etc.
  - ***Quantitative***: This is an ordinal dataset with a well-defined capability to perform arithmetic and comparison.
    - ***Example***: weight, height, scores, etc.

Direction of Ordering: 1/2

- **An order dataset can be *sequential* or *diverging*.**
- **A *sequential* dataset has a homogeneous range from a minimum to a maximum.**
- **A *diverging* dataset has data measured from a based point and extends to both ends. An elevation dataset is diverging because its measurement start at sea level.**

Direction of Ordering: 2/2

- **Both ends of a *diverging* dataset are sequential.**
- **A *cyclic* dataset has its values wrap around back to a starting point (e.g., the hour of the day, the day of a week, and month of the year, angle measures, etc.).**

Hierarchical Attributes

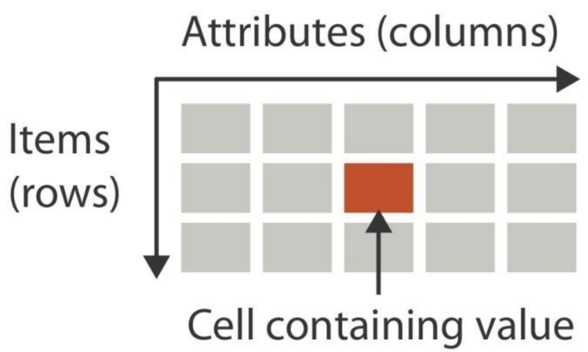
- **An attribute or across multiple attributes may have a hierarchical structure.**
- ***Example*: The daily prices of companies collected over the course of a decade is a *time-series* dataset.**
  - **One attribute is time. Moreover, time can be aggregated hierarchically (e.g., weeks, months, years). Each different aggregation may show interesting patterns.**
- **Geographic attribute of a postal code may be aggregated to the level of cities, states, etc.**

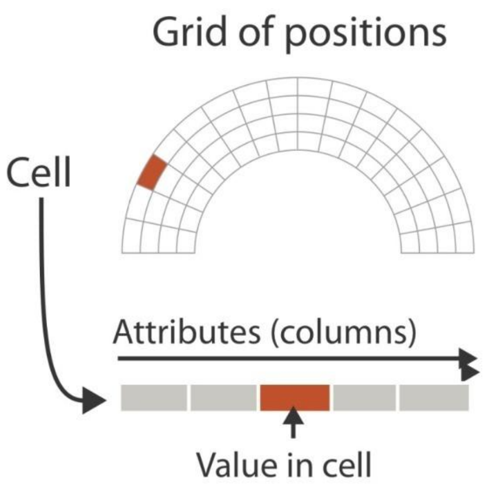
Data Semantics

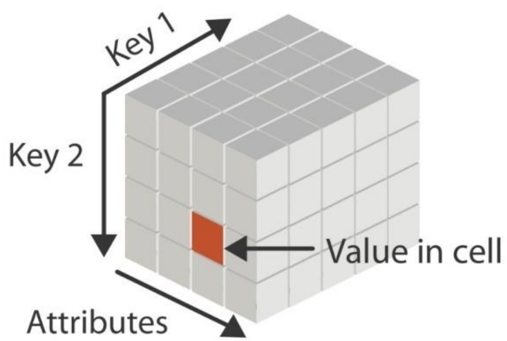
- **The *semantics* of the data is its real-world meaning.**
- **What does this number 9400109699937278441167 mean? Is it just a big integer? Is it a USPS tracking number or something else? Is it the number of days since the big bang?**
- **Is "Johnson" a person's name, a company's name, a city name, a password, a program name?**

Key vs. Value Semantics: 1/3

- A ***key*** attribute is an index used to find ***value*** attributes.







Key vs. Value Semantics: 2/3

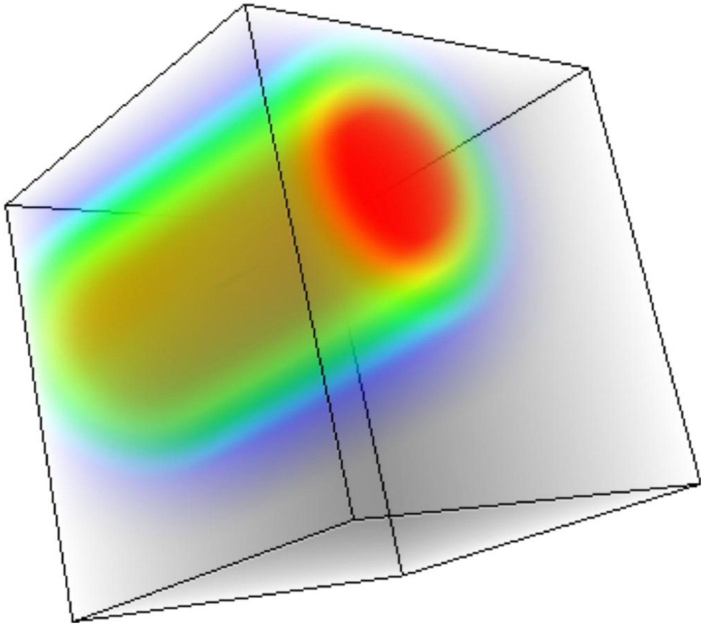
- **In a table, the key may just be implicit (e.g., the indices).**
- **A key attribute can be explicitly given. However, this key attribute must not have duplicate values.**
- **Recall this: Fields are generated through a systematic sampling so that each grid cell is a spanned by a unique range from a continuous domain.**

Key vs. Value Semantics: 3/3

- **Fields are *multivariate* (resp., *univariate*) if it has *more than* (resp. *only*) one value attributes.**
- **The *multidimensional* structure of a field depends on the number of keys.**
- **A field can be multivariate and multidimensional at the same time if it has multiple values and multiple keys.**

Scalar Fields

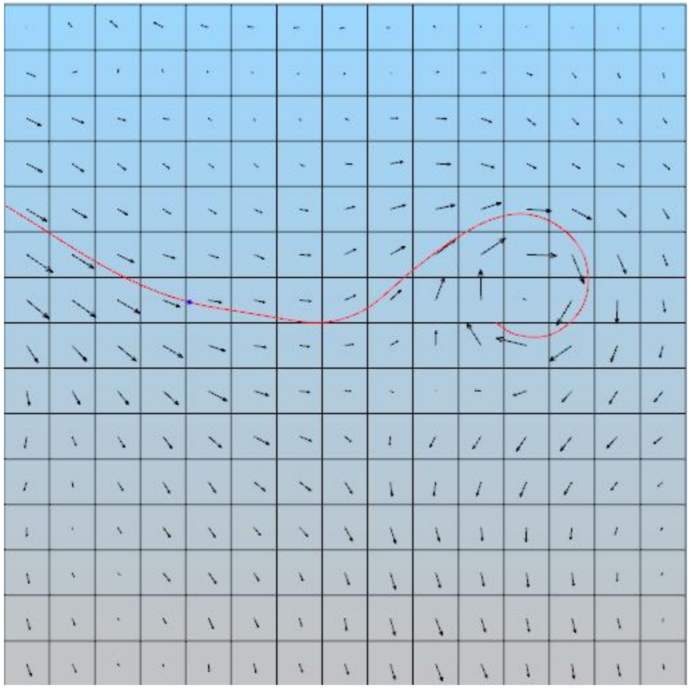
- A scalar field is ***univariate***, with a single attribute at each point in space.
- This means we assign a single value (e.g., temperature) to each point in space.



colors are used to show the values of a scalar field

Vector Fields

- A vector field is ***multivariate***, with a list of multiple attribute values (i.e., a vector) at each point in space.



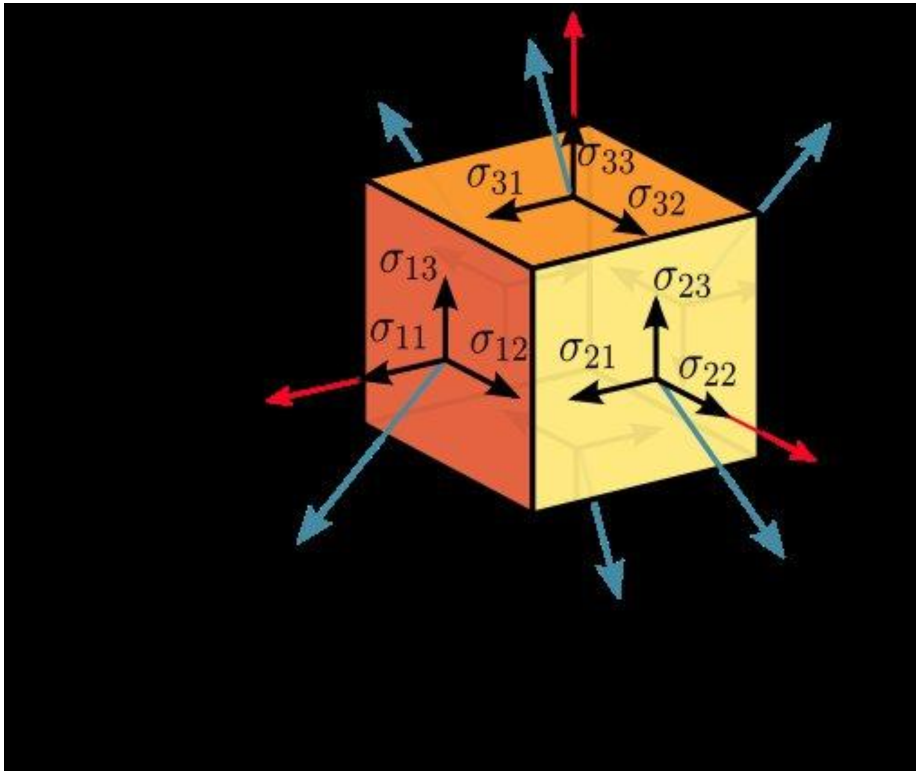
a vector is assigned to a representative point of a cell

Tensor Fields: 1/5

- **A tensor field is *multivariate* and has an array of attributes at a point. A rank n tensor in m-dimensional space is a mathematical object that has n indices and m^n components and obeys some transformation rules.**
- **Tensors are being used in differential geometry, physics, and engineering.**

Tensor Fields: 2/5

- ***Example***: **The Cauchy stress tensor or simply stress tensor that can be represented by a 3×3 real matrix $\boldsymbol{\sigma} = [\sigma_{ij}]_{3\times3}$.**



Wikipedia

Tensor Fields: 3/5

- ***Example*****: The metric measure on a manifold can also be represented as a tensor.**
- **The Euclidean metric tensor in the *n*-dimensional Euclidean space E^n is the $n \times n$ identity matrix.**
- **The curvature tensor measures the curvature at a point in a space.**

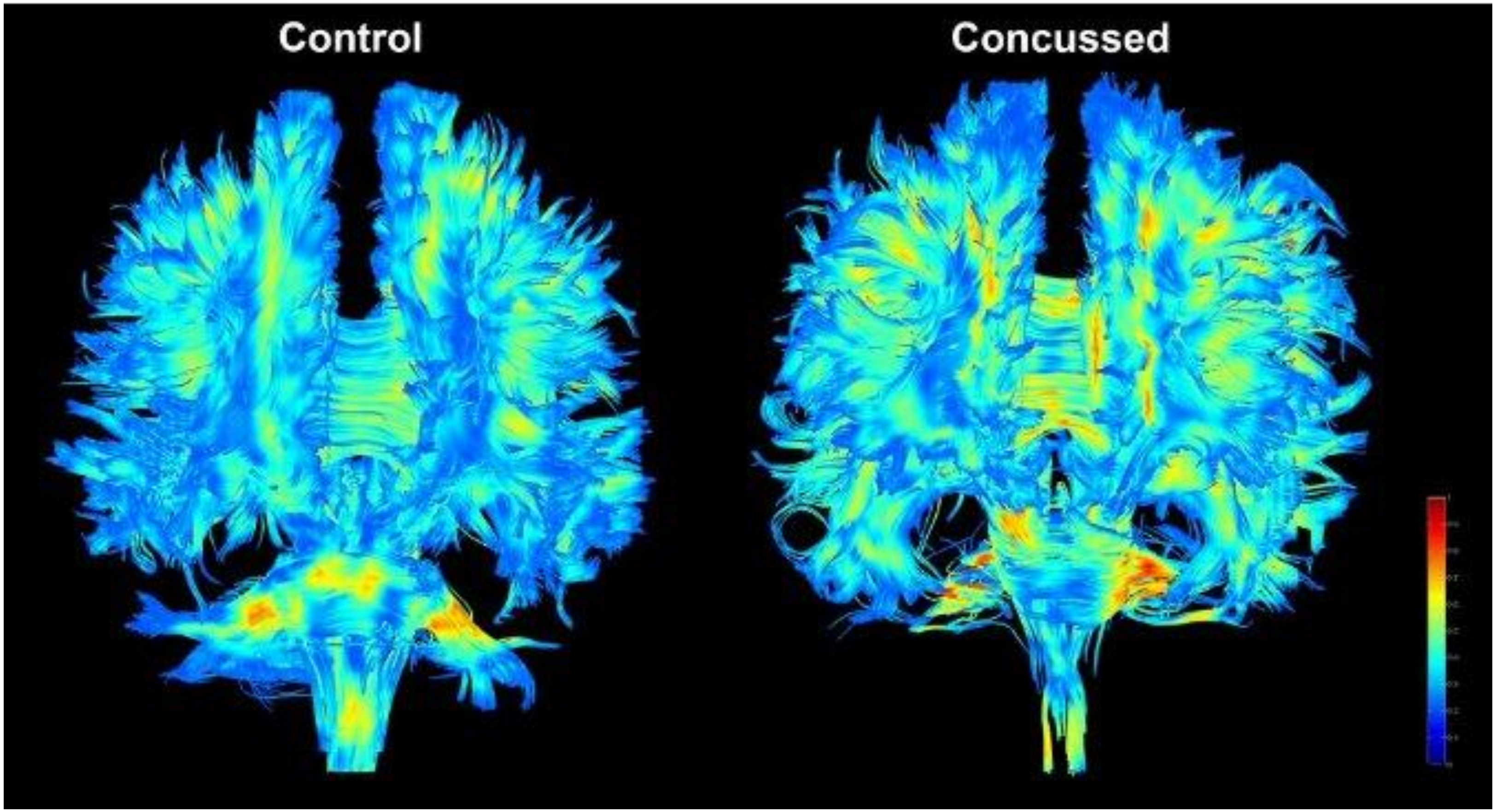
Tensor Fields: 4/5

- **The metric tensor of the Minkowski space, used in special relativity, is the one on the left.**
- **The Schwarzschild metric that describes the space-time around a spherical symmetric body is the one on the right.**

$$g = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix}$$

$$g_{\mu\nu} = \begin{bmatrix} (1 - \frac{2GM}{rc^2}) & 0 & 0 & 0 \\ 0 & -(1 - \frac{2GM}{rc^2})^{-1} & 0 & 0 \\ 0 & 0 & -r^2 & 0 \\ 0 & 0 & 0 & -r^2 \sin^2 \theta \end{bmatrix}$$

Tensor Fields: 5/5



Diffusion tensor visualization: The directionality of water diffusion. Warmer hues represent greater directional preference for water diffusion and cooler hues indicate more random water diffusion. – Michael Borich

Field Semantics

- **On previous slides, the categorization of spatial fields requires knowledge of the attribute semantics and cannot be determined from type information alone.**
- **Thus, if multiple measured values at each spatial point are given without further information, there is no sure way to know its structure.**

Temporal Semantics: 1/2

- **A *temporal* attribute is any kind of information that relates to time.**
- **Data about time is complicated to handle because of the rich hierarchical structure that can be used to reason about time and the potential for periodic structure.**
- **The analysis of time usually involves finding or verifying periodicity either at a predetermined scale or at some scale unknown in advance.**

Temporal Semantics: 2/2

- **It is important to note that there could be multiple ways to visually encoding that data.**
- **One of these ways may involve animation.**
- **A temporal key attribute can have either value of key semantics. For example,**
  - **The day/time (or duration) a transaction happened is a dependent value**
  - **Time can be an independent key – a MRI scan can have the independent keys of (*x*,*y*,*z*,*t*) to cover spatial position (*x*,*y*,*z*) and time *t*.**

Time-Varying Data: 1/2

- **A dataset has *time-varying* semantics when time is one of the key attributes, as opposed to when the temporal attribute is a value rather than a key.**
- ***Time-Varying Semantics*: The use of tracking device to track movements. The temporal attribute is an independent key.**
- ***Non Time-Varying Semantics*: A train scheduling table contains start time and end time. In this case, the time entries are values.**

Time-Varying Data: 2/2

- **A commonly seen temporal dataset is a *time-series* dataset.**
- **A time-series dataset usually has an ordered sequence of time-value pairs such as (*t*, *x*), where *t* and *x* are time and value.**
- **Note that the time values may not always be spaced at uniform temporal intervals.**
- **Typical time-series analysis tasks involve finding trends, correlations, autocorrelations, periods, variations at multiple time scales.**

The End