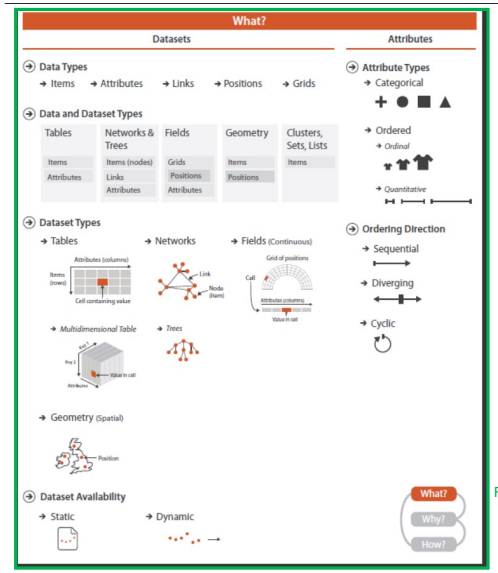
Data Types



From "Visualization Analysis & Design"
T. Munzner,
CRC Press, 2015
(Chapter 2)

Five different **data types**

- **item:** an object
- **link:** relationship between items
- **attribute:** property of an item
- **position:** a location in 2D or 3D space
- ***grid:*** *regular sampling of continuous data*

grid: more of an approach to collecting and storing data than a data type itself

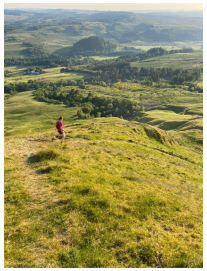
Running example

“Running” example

Hill running in Scotland

Runners take part in races

Races are held annually



Scottish Hill Racing: https://www.facebook.com/scottishhillracing/

Five different **data types**

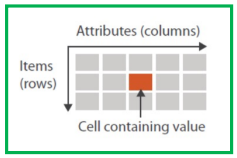
- **item:** a runner
- **link:** two runners train together (“run-buddies”)
- **attribute:** a runner belongs to a club
- **position:** the start point of a race
- **grid:** a runner’s heartbeat sampled every 30s

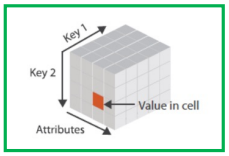
Four different **data set types**

- A data set type is a method for collecting data together
  - **table:** rows and columns (2D or multidimensional)
  - **networks and trees**: relationships between items
  - **fields:** continuous data *(conceptually there are an infinite number of measurements you could take, so sampling and/or extrapolation are necessary)*
  - **geometry:** spatial data

Data set type: table

Table: rows and columns (2D or multidimensional)

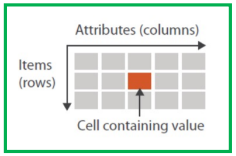


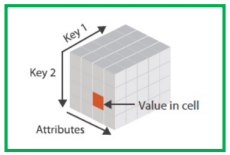


John Dunne	m	M50	Springly
Sara Ahmed	f	F60	Ludders
Mei Chan	f	F40	Bowlerside
Charles Ndlovu	m	M35	Ludders

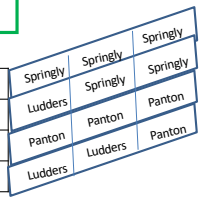
Data set type: table

Table: rows and columns (2D or multidimensional)





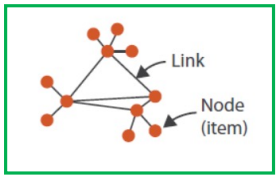
John Dunne	m	M50	Springly
Sara Ahmed	f	F30	Ludders
Mei Chan	f	F40	Bowlerside
Charles Ndlovu	m	M30	Ludders



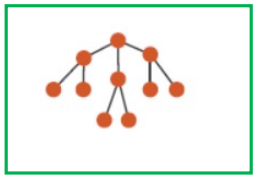
Modelling changes over time — e.g. adding in the 'key' or dimension of years, to represent changes in clubs (and age/gender categories) — makes the table multidimensional

Data set type: networks and trees

Networks and trees: relationships between objects



Links show run-buddies



Run-buddies are static pairs, but those pairs can group together in race events

Data set type: fields

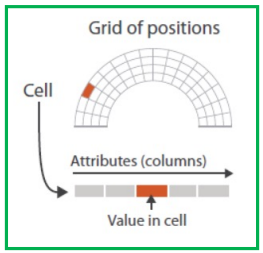
Fields: continuous data.

Conceptually there is an infinite number of measurements you could take, so sampling and extrapolation are necessary

every 10s

90bpm
100bpm
105bpm
106bpm
110bpm
115bpm
135bpm
140bpm

10s	90bpm
25s	102bpm
26s	103bpm
40s	106bpm
52s	112bpm
58s	114bpm
73s	137bpm
80s	140bpm



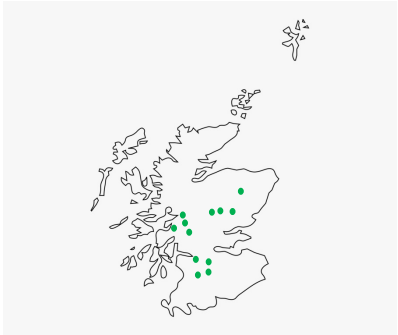
A runner's heartbeat. If you do not know the frequency of the heartbeat, e.g. every 10 seconds (as in the left example), then it has to be an attribute of the data item. Then you might use a table with two columns to represent the measurements as well as the times they were taken, as in the central example here.

The example on the bottom right is from Munzner's book, and I think it's to show a polar coordinate style sampling of a continuous region of space, i.e. positions are based on four distances from a central point, and an arc of 16 angles around that central point, so that the region has 64 samples... and so 64 cells.

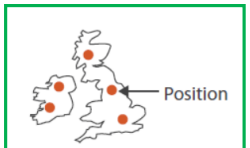
Each cell has 5 values measured or calculated for it, and we can see one in the bottom of the figure. One value in that cell is highlighted. It's unfortunate that the same colour used to highlight the value in the cell is also used to highlight the cell in the field. I would guess that Munzner is using this arc shape just to show that fields don't always have to be based on rectangular shapes or orthogonal axes. It's perhaps also unfortunate that the book doesn't explain this.

Data set type: geometry

Geometry: spatial data



Location of the annual Hill Running Races in Scotland, by start point



Note that if the *name* of a location might not be enough to be spatial data, partly as the name might be ambiguous. Also, names such as locations of hill races would have to be mapped on to a coordinate system, such as latitude and longitude (and perhaps altitude too), before we could call it spatial data.

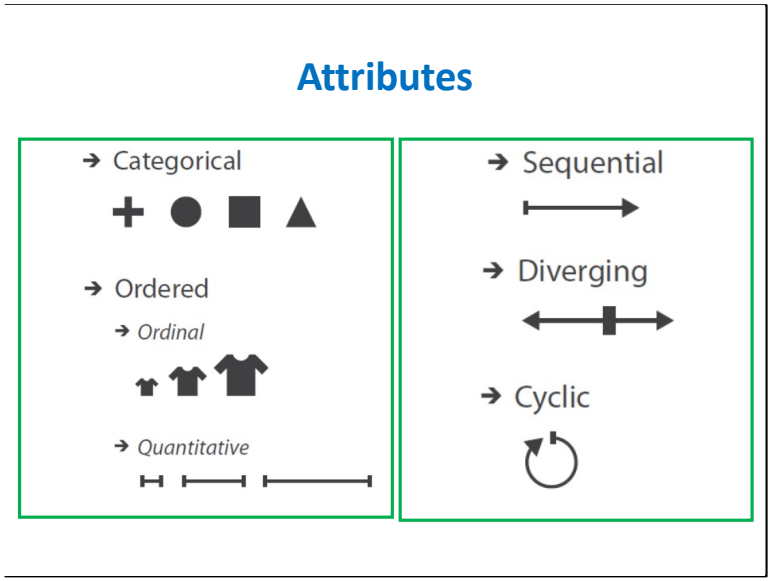
Data **Availability**

- Data is available at the same time, or collected as as dynamic stream
- Not the same as ‘data with a time dimension’
- ‘Online’ or ‘Offline’

Average finish time for the Two Breweries race in 2018

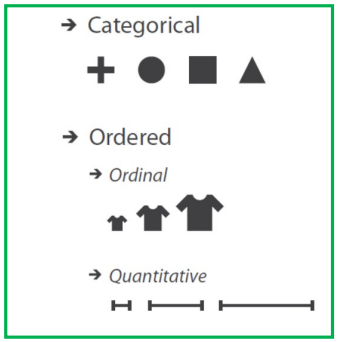
Average finish time for the Two Breweries race over all time

The assumption is that usually the data would all be available at the same time; streaming gives its own challenges



Attribute types and Ordering direction

Attribute Types



club: Springly, Ludders, Bolderside, Sharpford.

race difficulty:
- very difficult
- difficult
- managable by most runners
- easy
- very easy

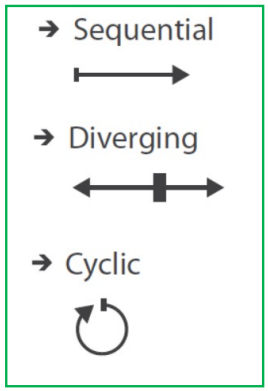
finishers' time: 1h40, 1hr42, 1hr53, 1hr54, 1hr58…

race date: 10th April, 15th Apr, 3rd May…

Categorical – also called 'nominal' – no implicit ordering, can only say whether things are the same or different.
Can impose an ordering (e.g. alphabetic, the number of members in a club), but this is imposed, not intrinsic

Ordinal has an intrinsic order, but the distances between items are not determined… so they cannot be added or subtracted
Quantitative involves a metric space, and so they can be added, subtracted (and – usually but not always – divided).

Ordering direction



Finisher's time for a race:
1h40, 1h42, 1h53, 1h54, 1h58…

elevation:
- 100m below
- 50m below
- 50m above
- 100m above

race date:
- 10th April
- 15th Apr
- 3rd May…
- …11th December
- 8th April
- 10th April

Race date is cyclic data…. although, of course, if you included the year in your Race Dates, it will be sequential data

Running example:
Two Breweries Hill Race (TBHR)

- Year
- Position
- Bib number
- Name
- Club
- Age category
- Finishing time

Running example:
Two Breweries Hill Race (TBHR)

Year, Position, Bib number, Name, Club, Age category, Finish time

1984	1	69	J Maitland	Aberdeen ACC	MOPEN	2:44:36
1984	2	53	B Brinfle	Horwich RMI	M50	2:50:36
1984	3	64	ARJ Curtis	Livingston & D	W50	2:52:34
1984	4	24	S Moore	Horwich RMI	M40	2:53:01
1984	5	65	AW Spenceley	Carnethy HR	WOPEN	2:56:55
1984	6	77	M Lindsay	Carnethy HR	MOPEN	2:58:42

I will leave it up to you to determine the data categories for each column

Running example:
Two Breweries Hill Race (TBHR)

Year, Position, Bib number, Name, Club, Age category, Finish time

1984	1	69	J Maitland	Aberdeen ACC	MOPEN	2:44:36
1984	2	53	B Brinfle	Horwich RMI	M50	2:50:36
1984	3	64	ARJ Curtis	Livingston & D	W50	2:52:34
1984	4	24	S Moore	Horwich RMI	M40	2:53:01
1984	5	65	AW Spenceley	Carnethy HR	WOPEN	2:56:55
1984	6	77	M Lindsay	Carnethy HR	MOPEN	2:58:42

- data types
- data set type
- data availability
- attribute types
- ordering direction

Summary

- **Data types**: nature of the data (5)
  - items, attributes, links, positions, grids
- **Data set types**: how the data is arranged (4)
  - tables, networks, fields, geometry
- **When the data is available** (2)
  - static, dynamic
- **Attributes**: properties of the data (2)
  - categorical, ordered (ordinal, quantitative)
- **Direction**: ways of ordering (3)
  - sequential, diverging, cyclic

Data Types