SPATIAL DATA STUDIO

Introduction

PhD Evelyn Uuemaa | PhD Alexander Kmoch | Bruno Montibeller

„Without geography, you are nowhere“, Jimmy Buffett

DR. EVELYN UUEMAA

Assoc. Professor in Geoinformatics

Education

1998–2003 BSc in physical geography and landscape ecology (*cum laude*); fulfilled also geoinformatics curriculum
2003–2004 diploma of geography teacher
2003–2004 MSc in geoinformatics and cartography
2004–2007 PhD in landscape ecology and environmental protection

Career

2003–2003 Ltd EOMap, cartographer
2007–2012 Researcher in Geoinformatics, Dep. of Geography, University of Tartu
2015–2017 EU Horizon 2020 Marie Curie fellow at the National Institute of Water and Atmospheric Research, New Zealand
2013–... Assoc. Professor in Geoinformatics, Dep. of Geography, University of Tartu
2020 –... Head of the Department of Geography

Research interests: spatial analysis, landscape analysis, geospatial land and water management, land use/cover changes, environmental modelling, water quality modelling, spatial machine learning

Workgroup lead: Landscape Geoinformatics
Twitter handle @evelynuuemaa

COURSE OUTLINE

- Geospatial data models, queries and coordinate systems
- Metadata, data quality, data formats, interoperability
- Introduction to remote sensing
- Data management, geoprocessing
- Visualisation, cartographic design
- Mid-term test
- Final project - The purpose of the final project is to synthesise the information learned in this course and demonstrate your skill in processing and handling spatial data and visualisation.

GRADING

- Perform all the individual assignments, self check tests and test. Each week usually gives 10 points. Once the task is being graded, it is **not possible to resubmit**. Possibility to earn some extra credits with additional tasks.
- Create a webpage for ePortfolio with one work included from Spatial Data Studio.
- Test – 20% of your grade
- Final Project – 20% of your grade
- Overall score will be a sum of individual assignments, test and final project.

EPORTFOLIO

- A website that enables you to collate digital evidence of your learning
- ePortfolios contain a wide range of digital files, including but not limited to, PDF files, videos, images and links to other websites or online resources. However, in our case you should focus on maps, visualisations, geodatabases, geoportals etc.
- Primarily a way to demonstrate (showcase) the highlights of a student's academic career

WHY WOULD YOU WANT AN EPORTFOLIO?

- ePortfolio is a creative, flexible, and powerful tool for applying a job, PhD position, starting your own business, or becoming a freelancer.
- ePortfolio is useful for you to show the employer that you as a job applicant have the knowledge and skills to succeed in their company.
- ePortfolio is a part of creating your own Digital Identity.
- ePortfolio is also recognised learning tool, and it enables you to reflect upon your strengths and weaknesses, and strive to improve ☺

COURSE WORKS THAT SHOULD BE INCLUDED IN EPORTFOLIO

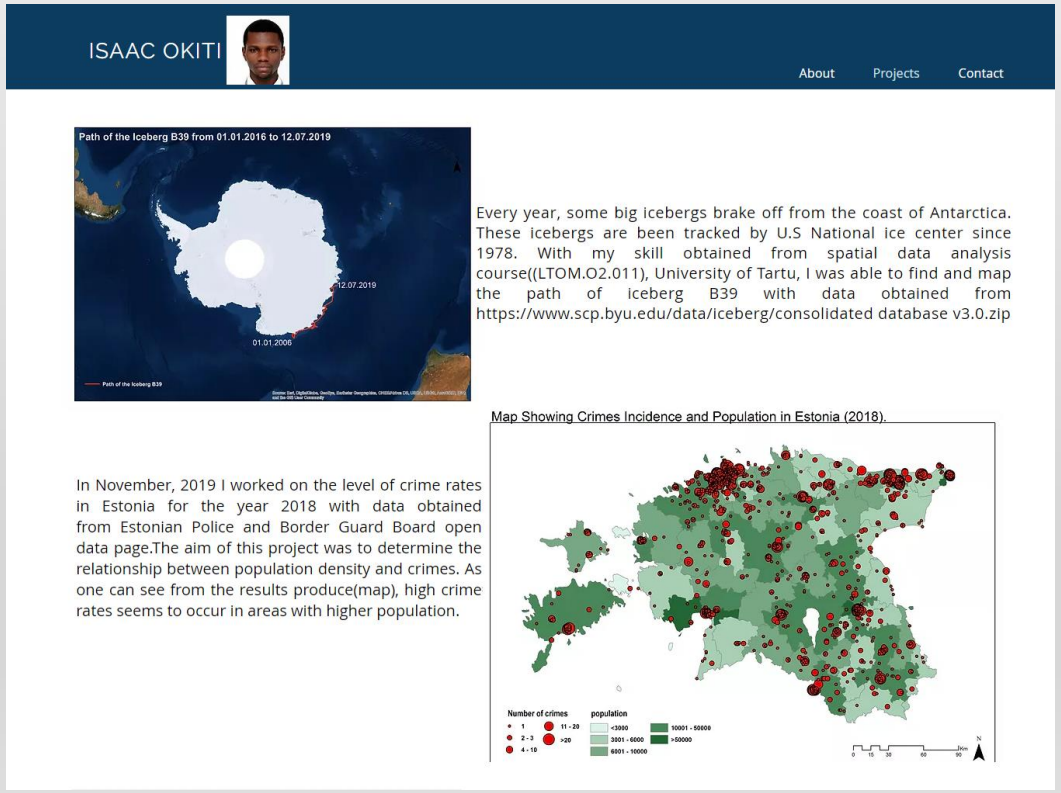
- **One work from Spatial Data Studio – COMPULSORY!**
- One work from either Planning Project, Spatial Data Analysis, Geography, Communication and Spatial Mobility, Visual Geodata Mining, Spatial Databases, Data Science in Remote Sensing, Work Placement, Demography and Urban Social Geography, 3D Modelling. Altogether **4-5 works from all these courses**.
- There is **no special session/time allocated** for ePortfolio – it is something you have to do along the courses as individual work. However, you can always ask assistance/advice.

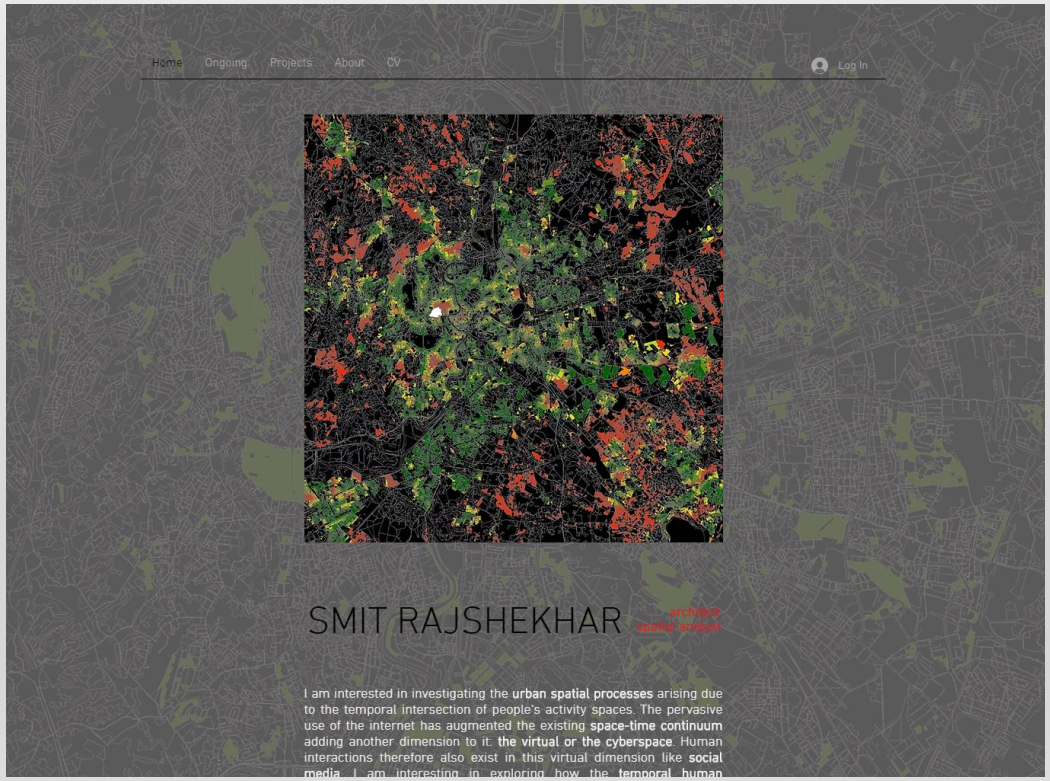
HOW TO PRESENT YOUR WORK IN EPORTFOLIO

- Add one or two of your nicest maps under each course (e.g. Spatial Data Studio) together with a short summary what was the aim of the project and how it was achieved, including programmes used
- There is no need to upload the whole report or show every little detail.
- Less is more! Make sure that your maps are meticulous and rather add less maps but perfect than more and not properly designed.
- Images, maps and photos in ePortfolio need to be **your** creation or otherwise you need to credit and be sure that the license allows them to use.

PREVIOUS YEAR EXAMPLES

- https://www.geograafia.ut.ee/en/studies/eportfolio





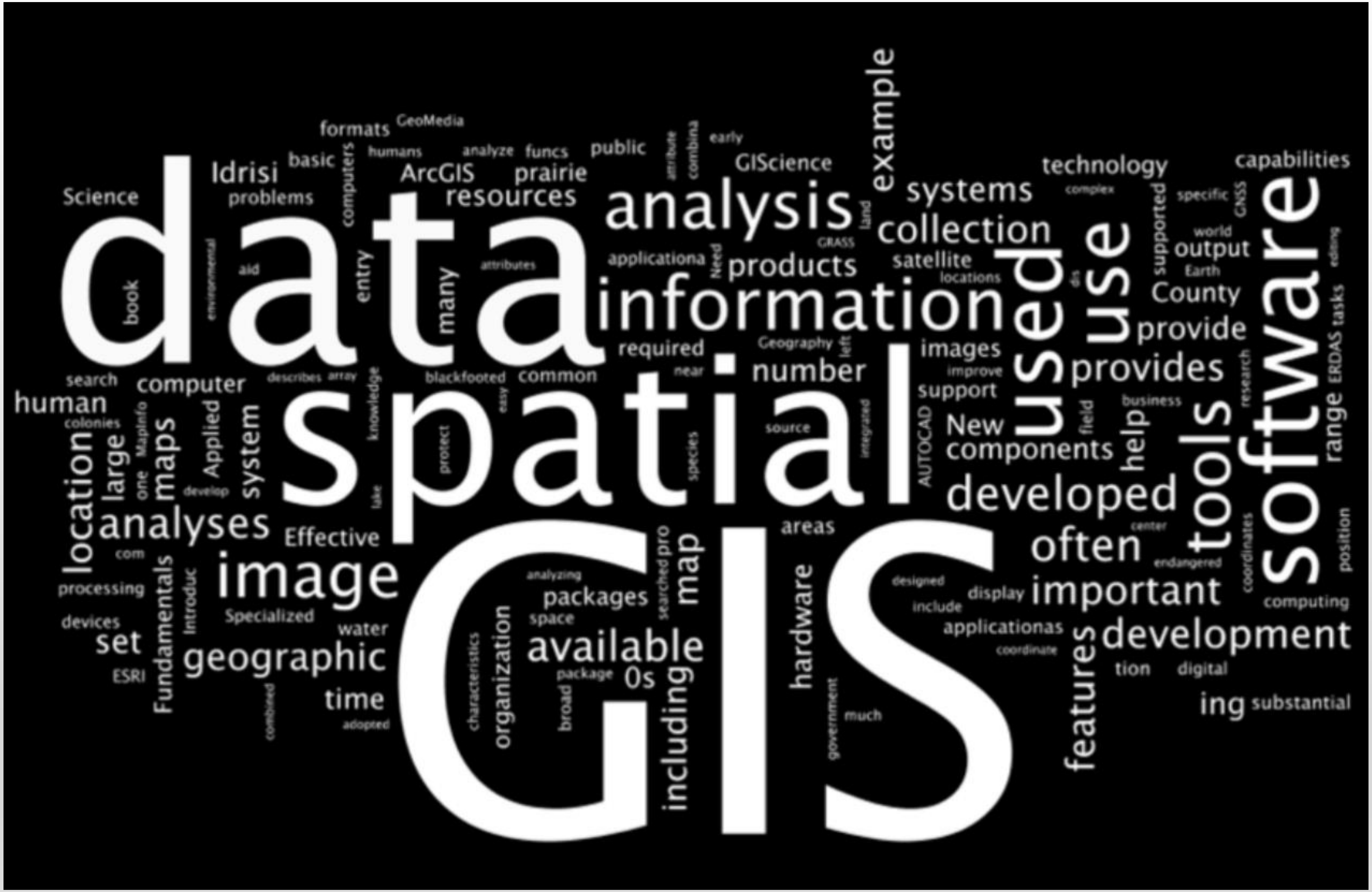
UNIVERSITY OF TARTU

Some free platforms for creating your ePortfolio

- sites.google.com
- https://www.youtube.com/watch?v=J9--8Pj2Gjs
- blogger.com
- https://www.youtube.com/watch?v=a6TX3eYqklo
- wordpress.com

ACADEMIC INTEGRITY

- Students are responsible for knowing the policy regarding academic honesty: Study Regulations of the University of Tartu
- Academic fraud *means in an assessment of learning outcomes, the use of any materials that the member of the teaching staff has not explicitly permitted the students to use; illicit sharing of knowledge (e.g. prompting, copying other student's work, etc.) by students participating in an assessment of learning outcomes; participating in an assessment of learning outcomes for another student; submission of the written work of another person as the student's own, or the use of parts thereof without the appropriate academic reference; second submission of the student's own work, if the student has already received ECTS for it.*


GIS
data
spatial
information
analysis
software
used
use
tools
image
analyses
geographic
available
developed
development
important
often
features
location
including
hardware
organization
collection
systems
products
example
technology
capabilities
resources
ArcGIS
Idrisi
prairie
Science
problems
basic
formats
GeoMedia
humans
analyze
funcs
public
GIScience
early
applicationa
satellite
output
County
provide
provides
images
number
required
Geography
computer
human
search
blackfooted
common
support
New
components
help
range
ERDAS
tasks
research
business
field
AUTOCAD
source
near
improve
species
maps
large
Applied
system
knowledge
book
entry
many
aid
attributes
Effective
areas
map
packages
space
processing
Fundamentals
Introduc
devices
Specialized
water
set
ESRI
time
0s
package
broad
characteristics
combined
adopted
government
much
feature
tion
digital
ing
substantial
computing
coordinates
position
endangered
center
designed
display
include
applicationas
coordinate
analyzing
searched
pro
com
colonies
MapInfo
one
develop
describes
array
protect
lake
environmental
attribute
combina
land
GRASS
Need
locations
left
world
Earth
specific
GNSS
supported
complex
editing
integrated
easy

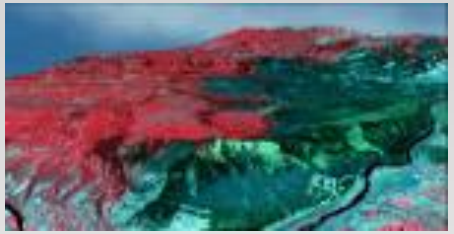
WHAT IS GIS?

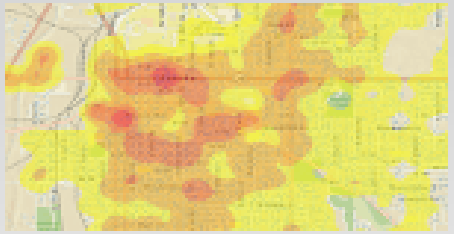
- A GIS is a computer-based system to aid in the collection, maintenance, storage, analysis, output, and distribution of spatial data and *non-spatial* data and information.

- GIS can ingest any type of data – both spatial and non-spatial
- We are currently in „data overload era“ – visualising data helps to understand data better
- **Half of all the time** spent on a GIS projects will simply be **working with data**

HOW CAN GIS BE USED?

- GIS in mapping
- Telecom and network services
- Urban planning
- Environmental Impact analysis
- Disaster management and mitigation
- Wildlife mapping

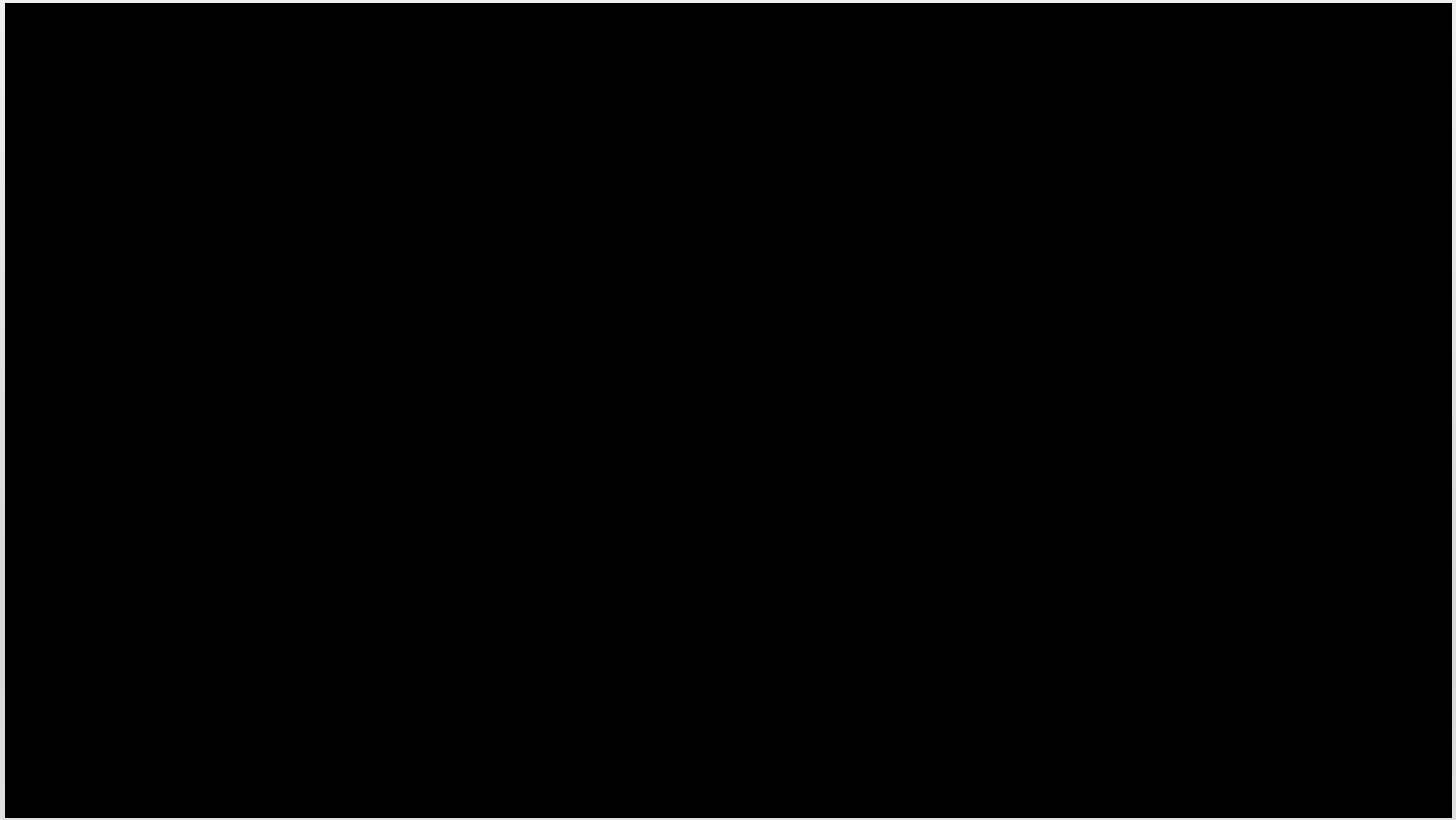




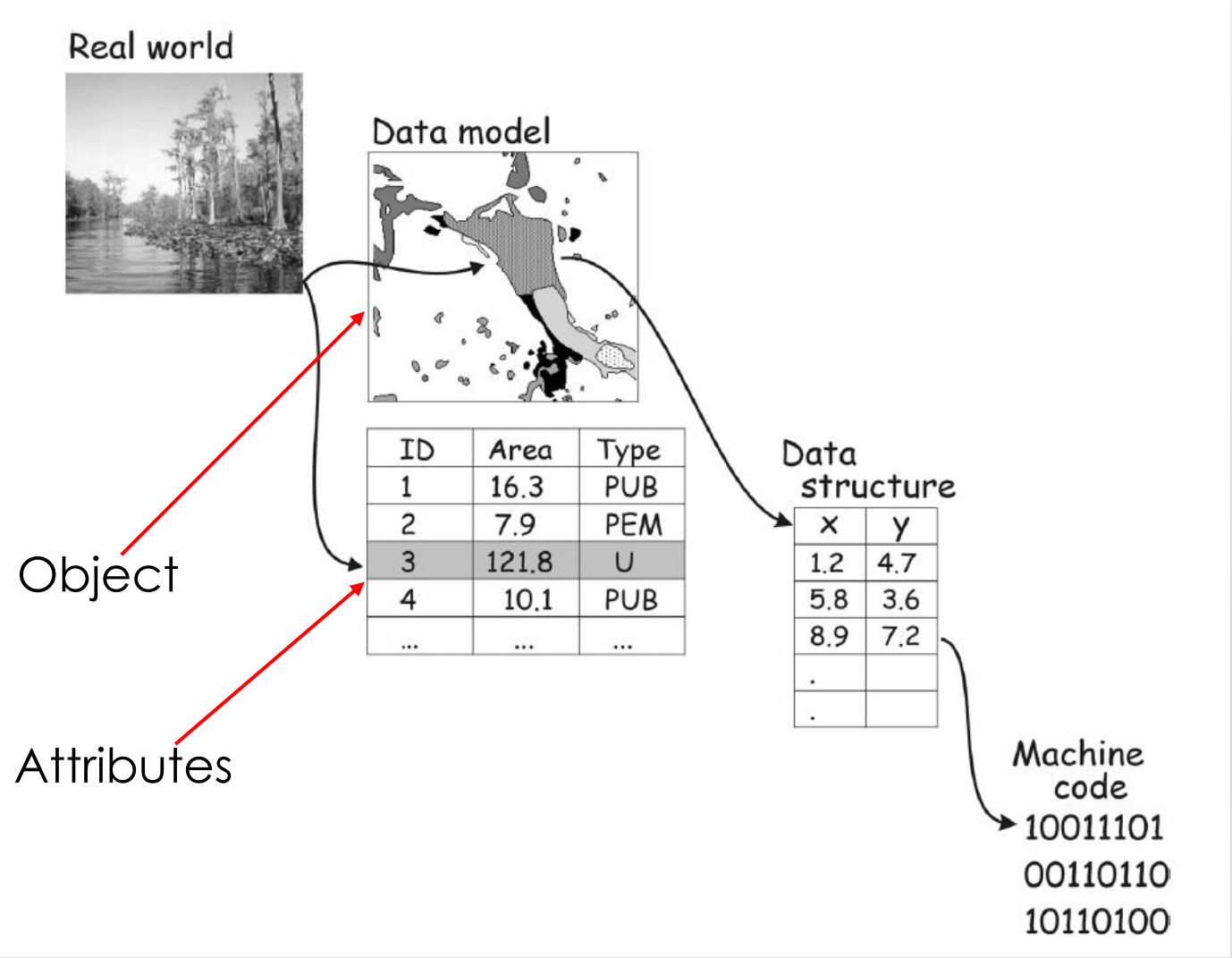
Source: http://grindgis.com/blog/gis-applications-uses

GEOSPATIAL TECHNOLOGY IS EVERYWHERE

https://www.youtube.com/watch?v=ZdQjc30YPOk



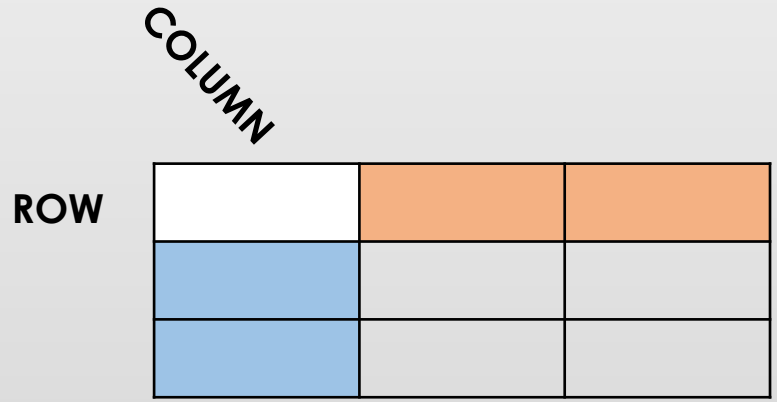
GEOSPATIAL DATA MODELS



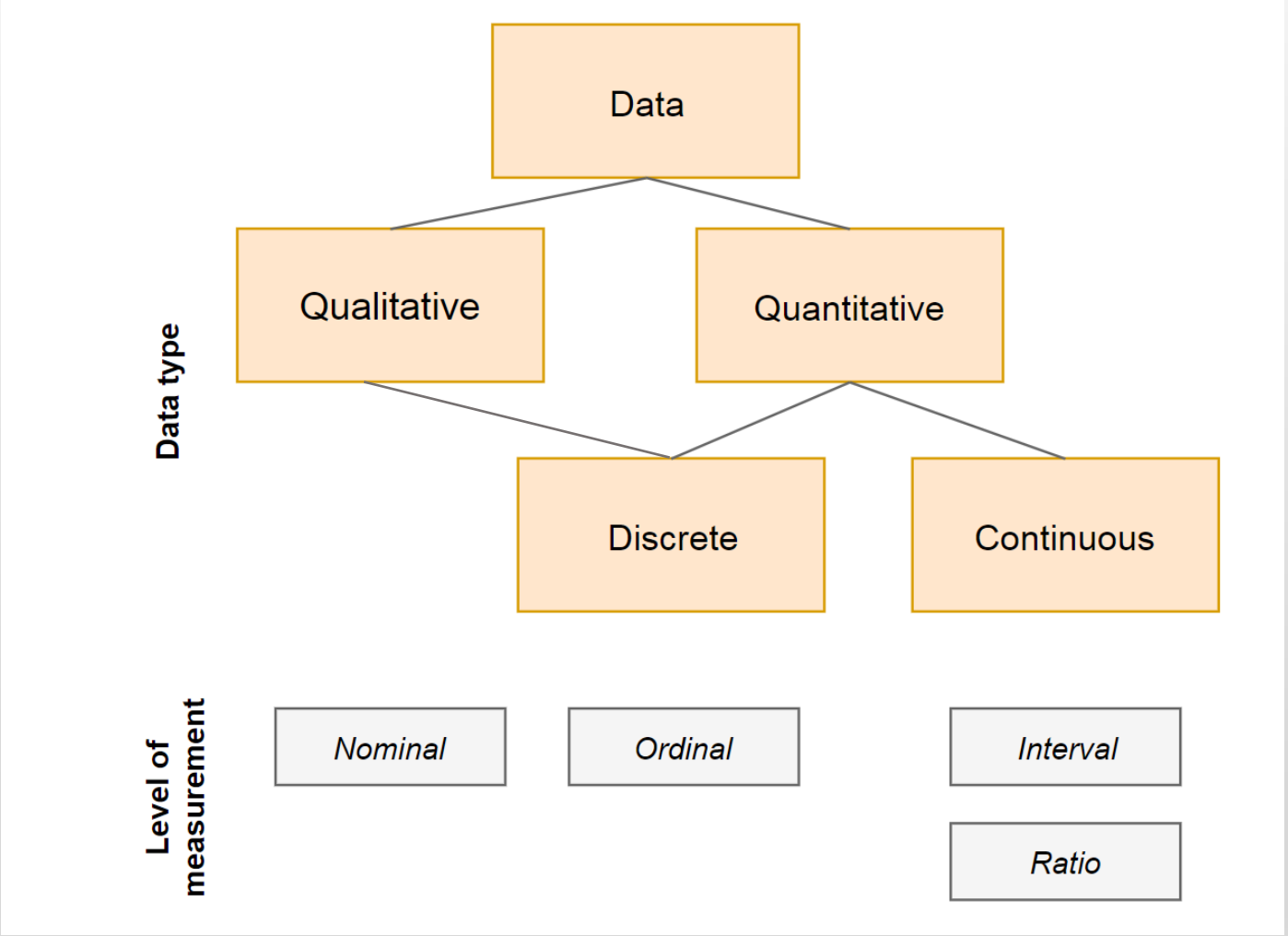
Source: Bolstad, 2008

ATTRIBUTES

- Record non-spatial characteristics that describe spatial entity
- Arranged in tables
  - ✓ Row = 1 entity
  - ✓ Column = 1 attribute
- Stored in a computer in a flat-file format or a Database Management System



ATTRIBUTE CATEGORIES - TYPES OF DATA AND LEVEL OF MEASUREMENT

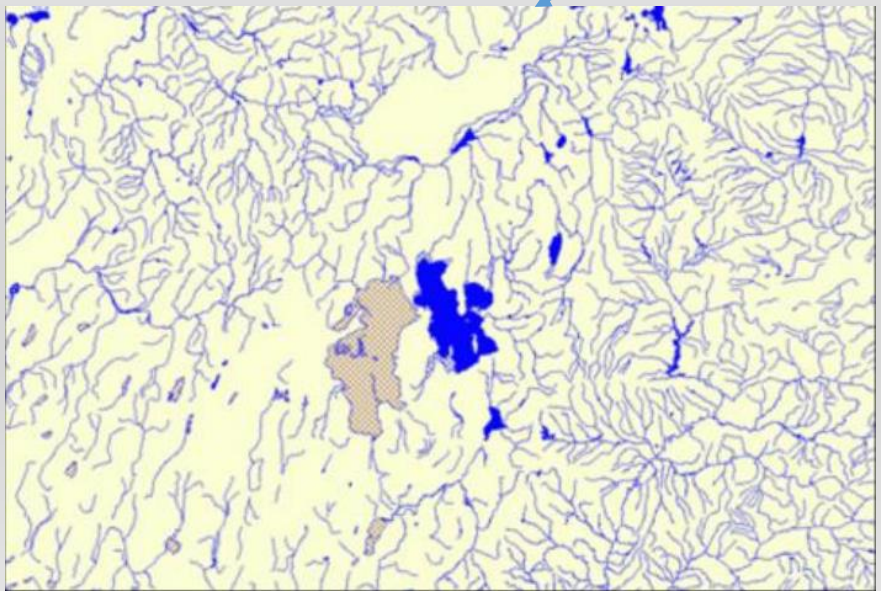


Stevens, 1946

Different mathematical operations on variables are possible, depending on the level at which a variable is measured (e.g. forest + urban area = ?)

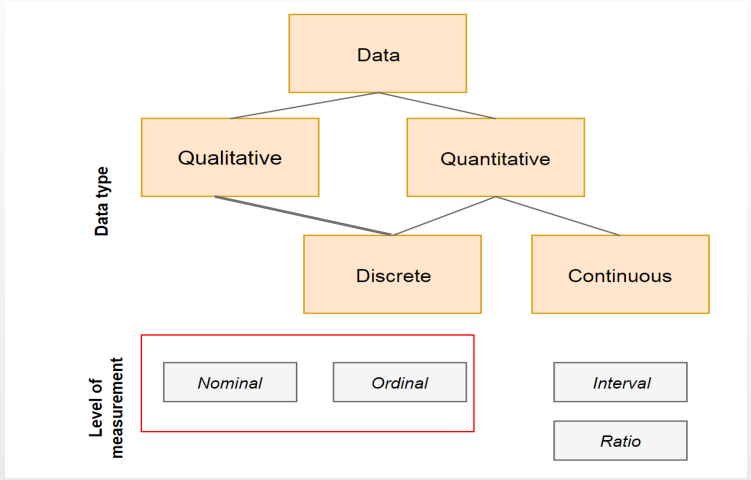
Discrete

- Individually distinguishable
- Does not exist between observations
- Examples: streams and lakes, roads



Continuous

- Exist between observations
- Represent data of a continuous nature
- Examples: temperature, elevation



Nominal

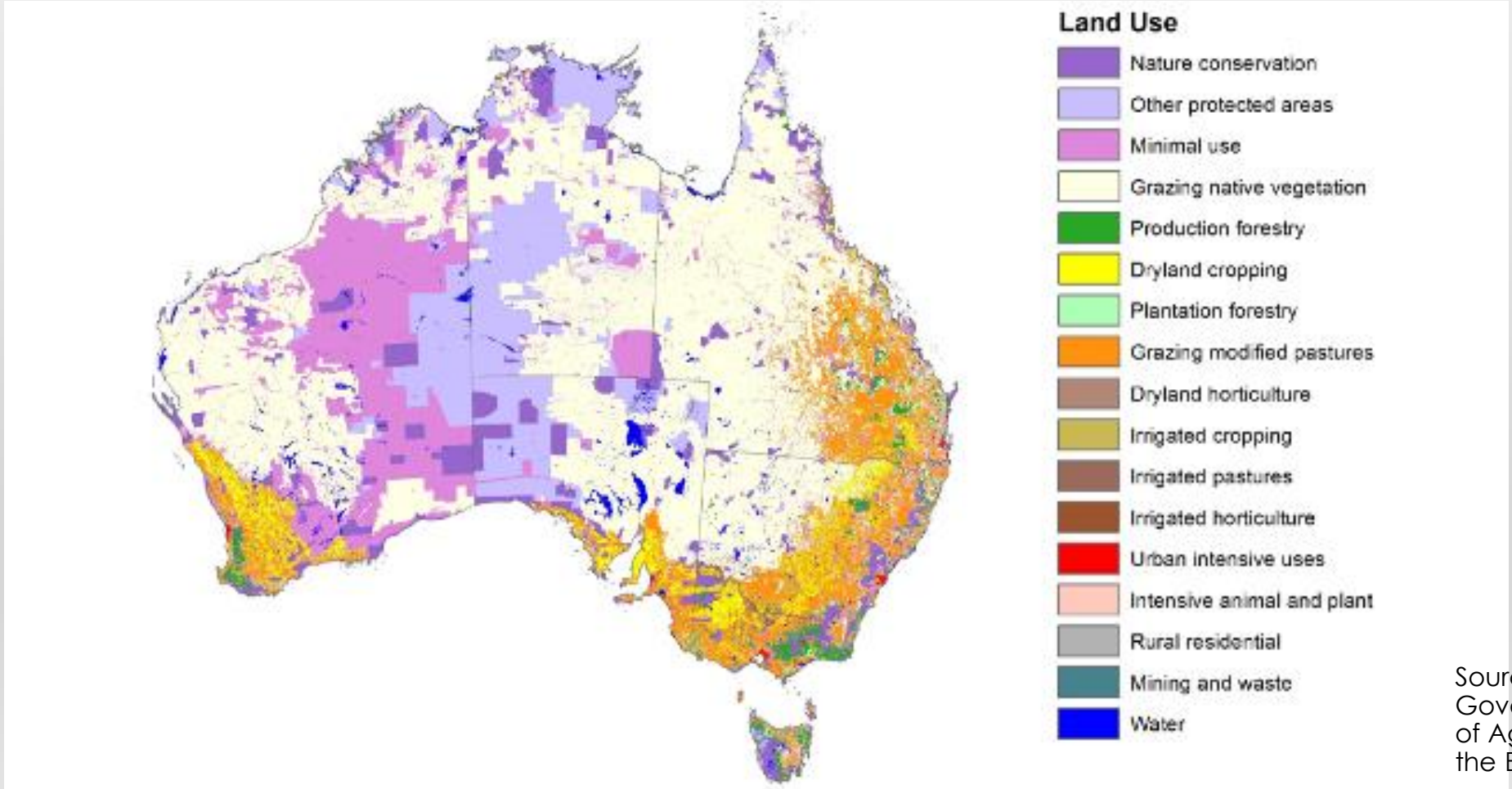
- Data categories are represented by labels or names
- Even if the labels are numerically coded, the data categories have **no logical order**
- Central tendency - mode
- Examples: land use, gender, religious affiliation

Ordinal

- Data classifications are represented by sets of labels or names (high, medium, low) that have relative values
- The classified data **can be ranked or ordered**
- Central tendency - median
- Examples: small, medium and large size coffee

NOMINAL MAP

- Land use of Australia

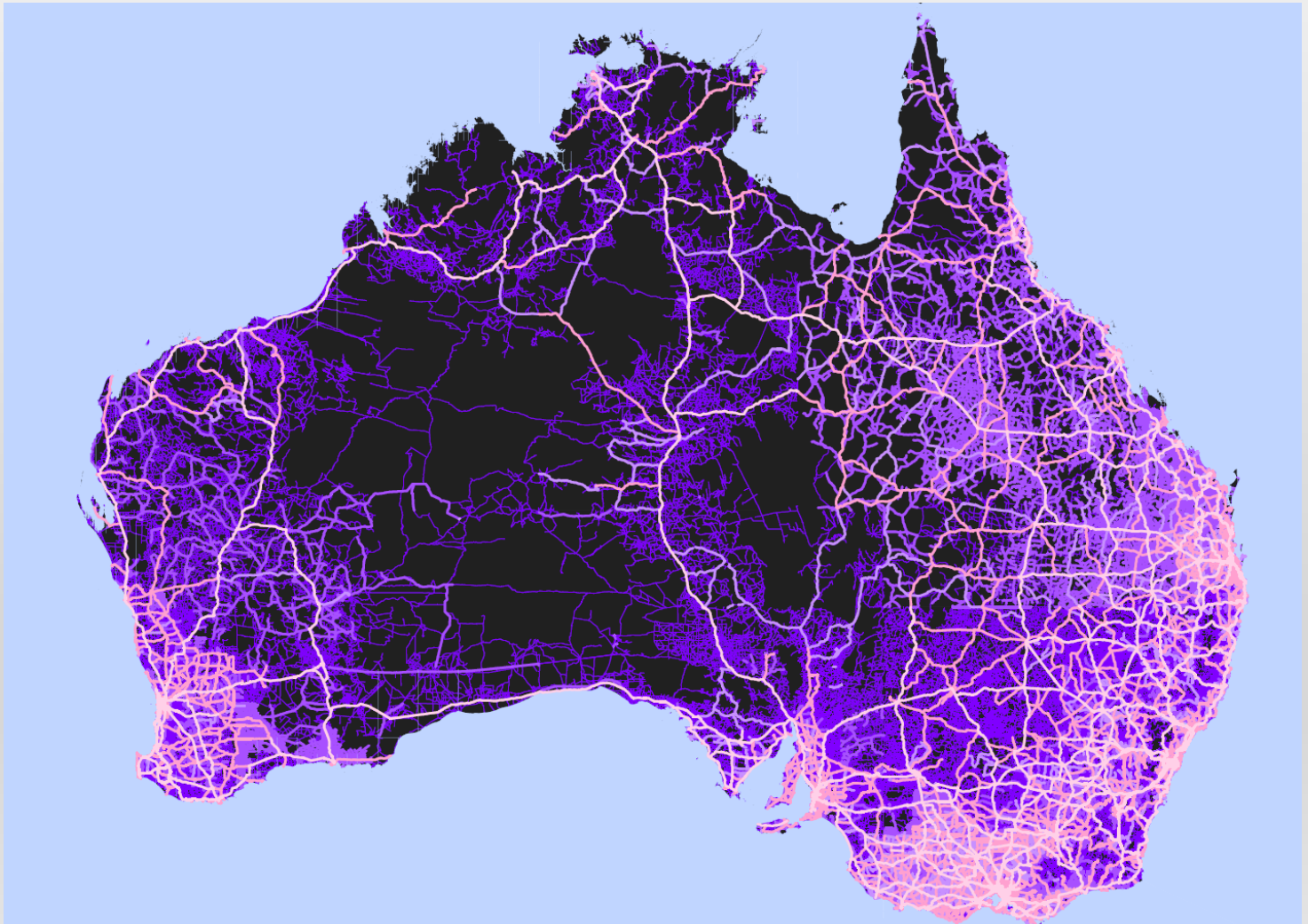


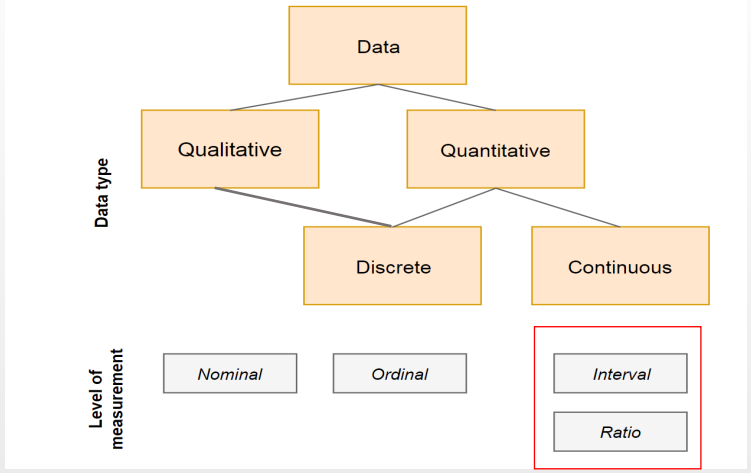
Source: Australian Government, Deparmentd of Agriculture, Water and the Environment

- Road map of Australia (National highways)

ORDINAL MAP

- Road map of Australia – roads colored by road type





Interval

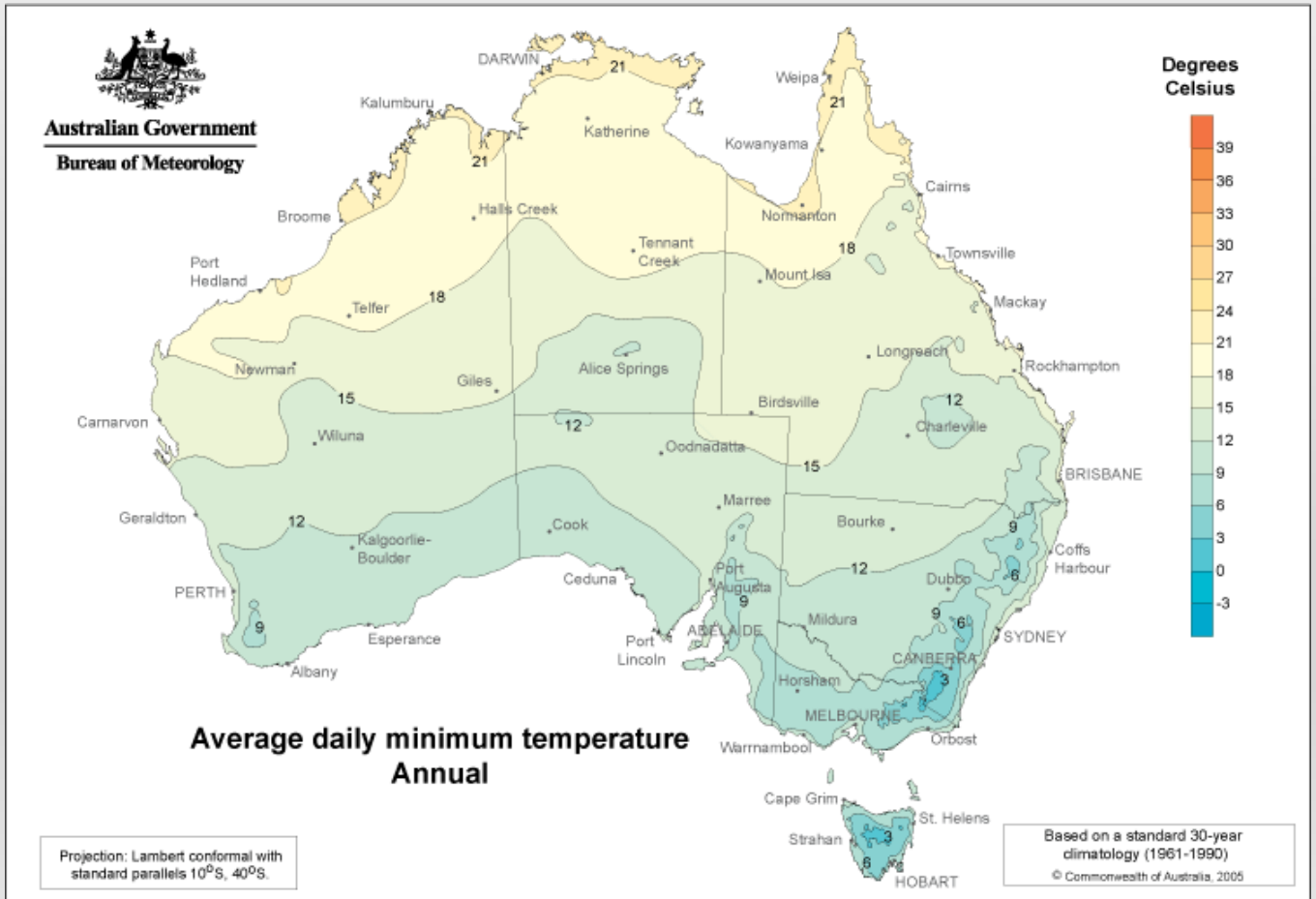
- Similar to the ordinal level with the additional property that meaningful amounts of differences between data values can be determined.
- There is no natural zero point.'
- Central tendency – mode, median, arithmetic mean
- Example – temperature on Fahrenheit scale

Ratio

- The interval level with an inherent zero starting point.
- Differences and ratios are meaningful for this level of measurement.
- Central tendency – mode, median, arithmetic mean, geometric mean
- Examples: monthly income, Kelvin scale

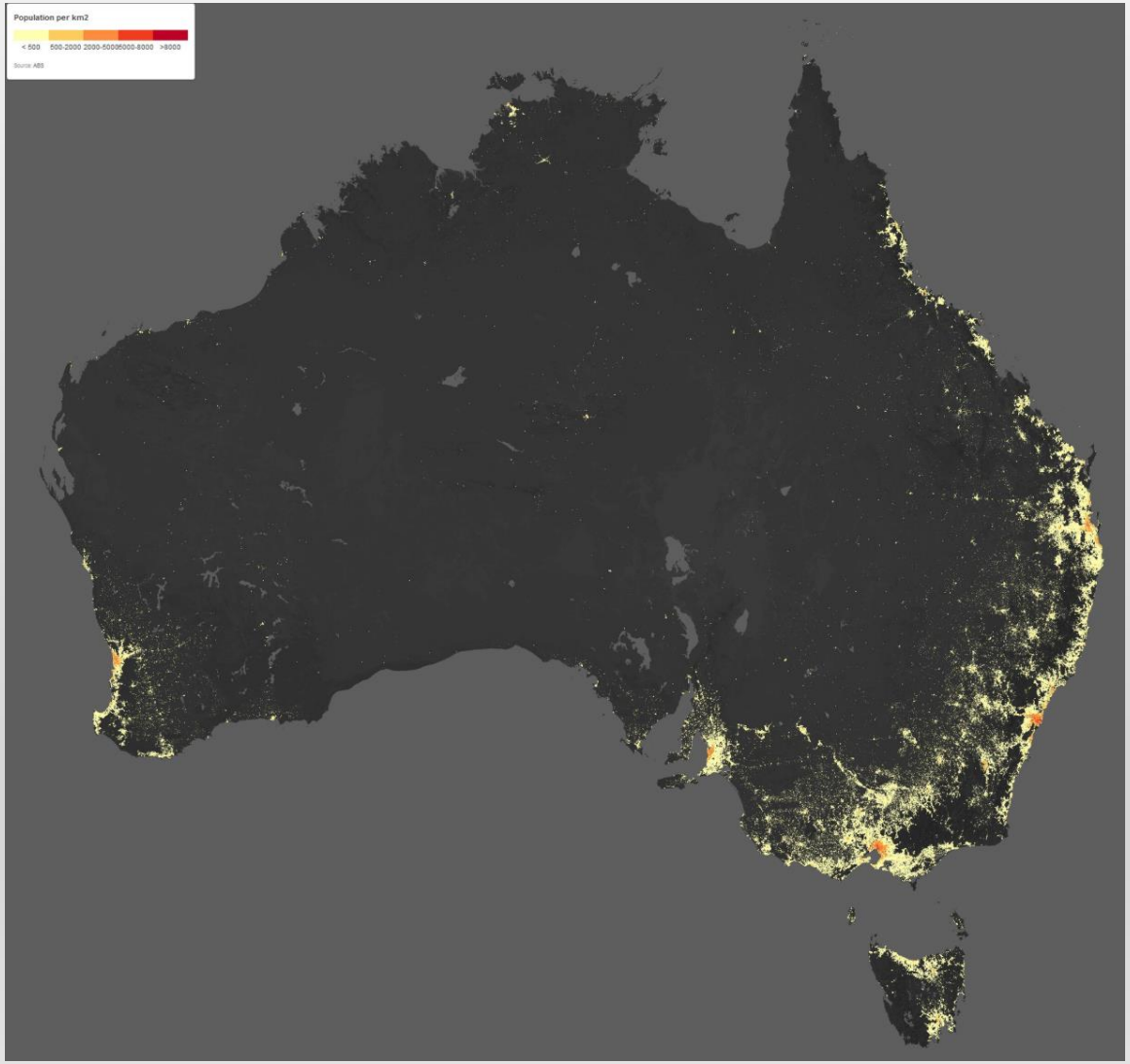
INTERVAL MAP

- Temperature map of Australia

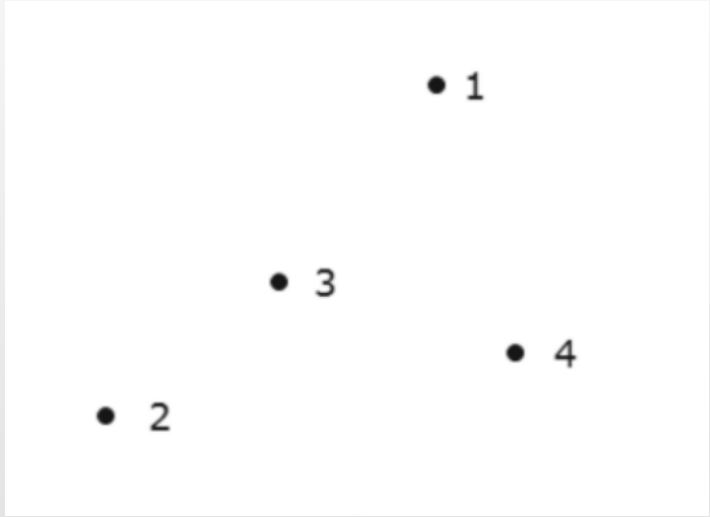


RATIO MAP

- Population density of Australia



Identify Columns as:
Nominal
Ordinal
Interval/Ratio



ID	Height (m)	Type	Class
1	15	Pine	Medium
2	20	Pine	Medum
3	10	Olive tree	Small
4	30	Oak	Tall

Attribute data types

- Attributes are stored in computer memory
- The data type of the attribute needs to be specified for **efficient use of memory** and determination of operation applicability
- There are four typical data types:
  - ✓ Integer
  - ✓ Float/Real
  - ✓ Text/String
  - ✓ Date

Integer

- e.g. Whole number
- Can be used for mathematical calculations

  However, any resulting fraction of a whole number will be rounded
- Examples:

✓ 2

✓ 345

✓ -78

Float/Real

- e.g. Decimal number
- Can be used for mathematical calculations
- Examples:
  - ✓ 1.54
  - ✓ 345.0988
  - ✓ -45.09

Text/String

- e.g. characters
- Cannot be used for mathematical calculations
- Strings can be manipulated, extract substrings
- Examples:
  - ✓ „a“
  - ✓ „tree“
  - ✓ „Vanemuise 46“
  - ✓ „13“

Date

- Holds date information
- Cannot be used for mathematical calculations
- Lengths of time can be calculated
- Examples:
  - ✓ 12/11/2018
  - ✓ 15.10.17
  - ✓ 4 August 2005
- QGIS and ArcGIS format the date as datetime yyyy-mm-dd hh:mm:ss AM or PM.

- ISO 8601 - International standard for representation of dates, times and duration
  - ✓ Uses the Gregorian calendar system
  - ✓ Ordered from most to least significant: year, month, day, hour, minute
  - ✓ Each date and time value has a fixed number of digits

Source: http://dinesql.blogspot.com

- Duration: P[n]Y[n]M[n]DT[n]H[n]M[n]S

PT8H30M30s

Bad example: „the measurement was made at 8 in the morning on 6 th of October, 2017 and it took 1 hour 30 minutes and 30 seconds“

CHOOSING THE DATA (FIELD) TYPE

- In choosing the data type, first consider if your data is date, name or number. If you have dates then your field type is Date. For names, addresses - String
- However, an alternative to using repeating textual attributes is to establish a coded value. A textual description would be coded with a numeric value. For example, you might code road types with numeric values by assigning a 1 to paved improved roads, a 2 to gravel roads, and so on. This has the advantage of using less storage space in the spatial database; however, the coded values must be understood by the data user.

- If you have numbers then deciding whether you need short integer, long integer, float or double is a bit more complicated.

- First consider the need for whole numbers versus fractional numbers. If you just need to store whole numbers, such as 12 or 12345, specify a short or long integer. If you need to store fractional numbers that have decimal places, such as 0.23 or 1234.5678, specify a float or a double.

- Secondly, when choosing between a short or long integer, or between a float or double, choose the data type that takes up the least storage space required. This will not only minimize the amount of storage required but will also improve performance. If you need to store integers between -32,768 and 32,767 only, specify the short integer data type, because it takes up only 16 bits, whereas the long integer data type takes up 32. If you need to store fractional numbers between -3.4E-38 and 1.2E38 only, specify the float data type, because it takes up 32 bytes, whereas the double data type takes up 64.

- Specifying the precision and scale allows you to restrict the range of values and number formats a field can accept, giving you greater control.

Data type (QGIS)	Data type ArcGIS	Storable range	Size (Bits)	Applications
Integer16	Short integer	-32,768 to 32,767	16	Numeric values without fractional values within specific range; coded values
Integer32	Long integer	-2,147,483,648 to 2,147,483,647	32	Numeric values without fractional values within specific range
Integer64			64	Numeric values without fractional values within specific range
Real (adjusts automatically based on the determined precision)	Float (single-precision floating-point number)	approximately -3.4E38 to 1.2E38	32	Numeric values with fractional values within specific range
Real (adjusts automatically based on the determined precision)	Double (double-precision floating-point number)	approximately -2.2E308 to 1.8E308	64	Numeric values with fractional values within specific range

CHARACTER ENCODING

	LU04GBPER209	kult4569	Herr&#65533;tizopf P5352	
	LU04GBPER209	kult4569	Herr&#65533;tizopf P5352	
	LU04GBPER209	kult4569	Herr&#65533;tizopf P5352	
	LU04GBPER209	kult4568	Nelkenstr 11 Um...	
	LU04GBPER209	kult4559	Lauerzring	
	LU04GBPER209	kult4559	Lauerzring	
	LU04GBPER209	kult4559	Lauerzring	
	LU04GBPER209	kult4630	Gartenh&#65533;tten L...	
	LU04GBPER209	kult4630	Gartenh&#65533;tten L...	
	LU04GBPER209	kult4630	Gartenh&#65533;tten L...	

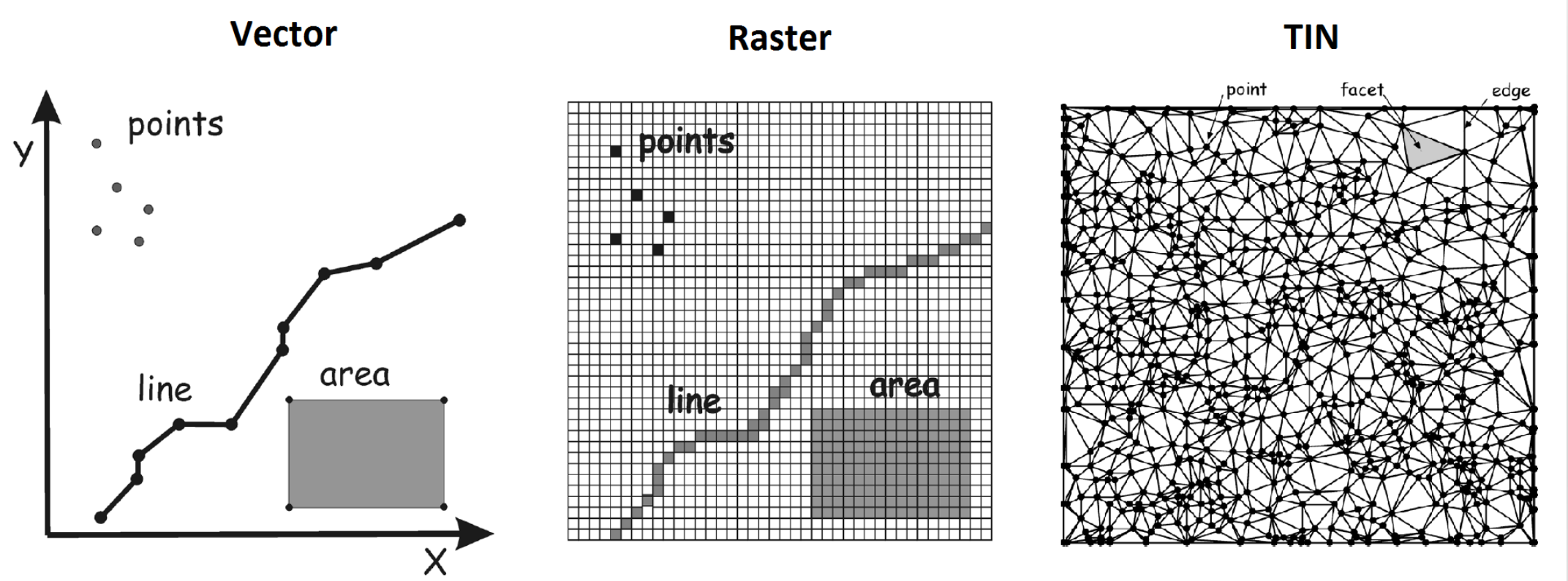
```
00:04:35,480 --> 00:04:37,010
ÄãÖªµÀ, ÎÒÃÇÕý½«·âÃæ·Åµ½...

51
00:04:37,050 --> 00:04:39,980
ËùÓÐµÄTPS±¨¸æÉÏ£¬ÔÚËûÃÇ³öÈ¥Ö®Ç°.
```

- Words and sentences in text are created from **characters**. Examples of characters include the Latin letter á or the Chinese ideograph 請
- Characters that are needed for a specific purpose are grouped into a **character set** (also called a **repertoire**). To refer to characters in an unambiguous way, each character is associated with a number, called a **code point**.
- The characters are stored in the computer as one or more **bytes**.
- A **character encoding** provides a key to unlock (ie. crack) the code. It is a set of mappings between the bytes in the computer and the characters in the character set. Without the key, the data looks like garbage.
- ArcGIS and QGIS use UTF-8 (UNICODE) by default.

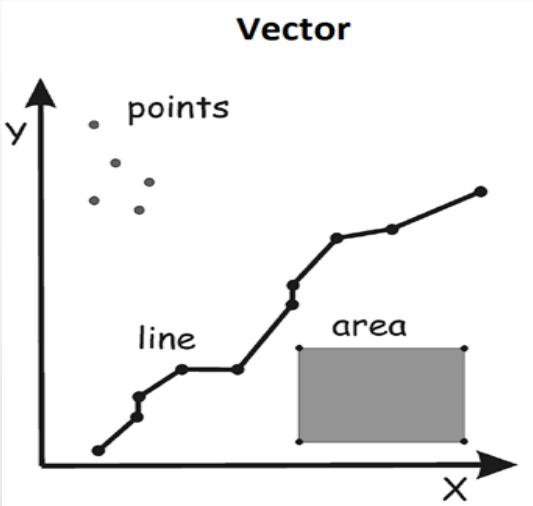
SPATIAL DATA MODELS II

There are three common spatial data models being used in GIS today:



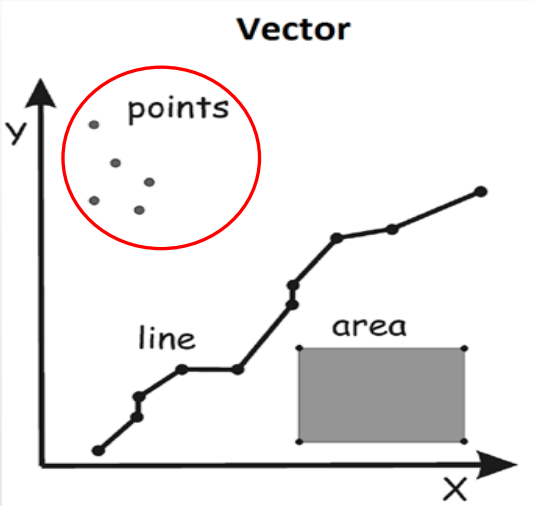
Source: Bolstad, 2008

VECTOR DATA MODEL



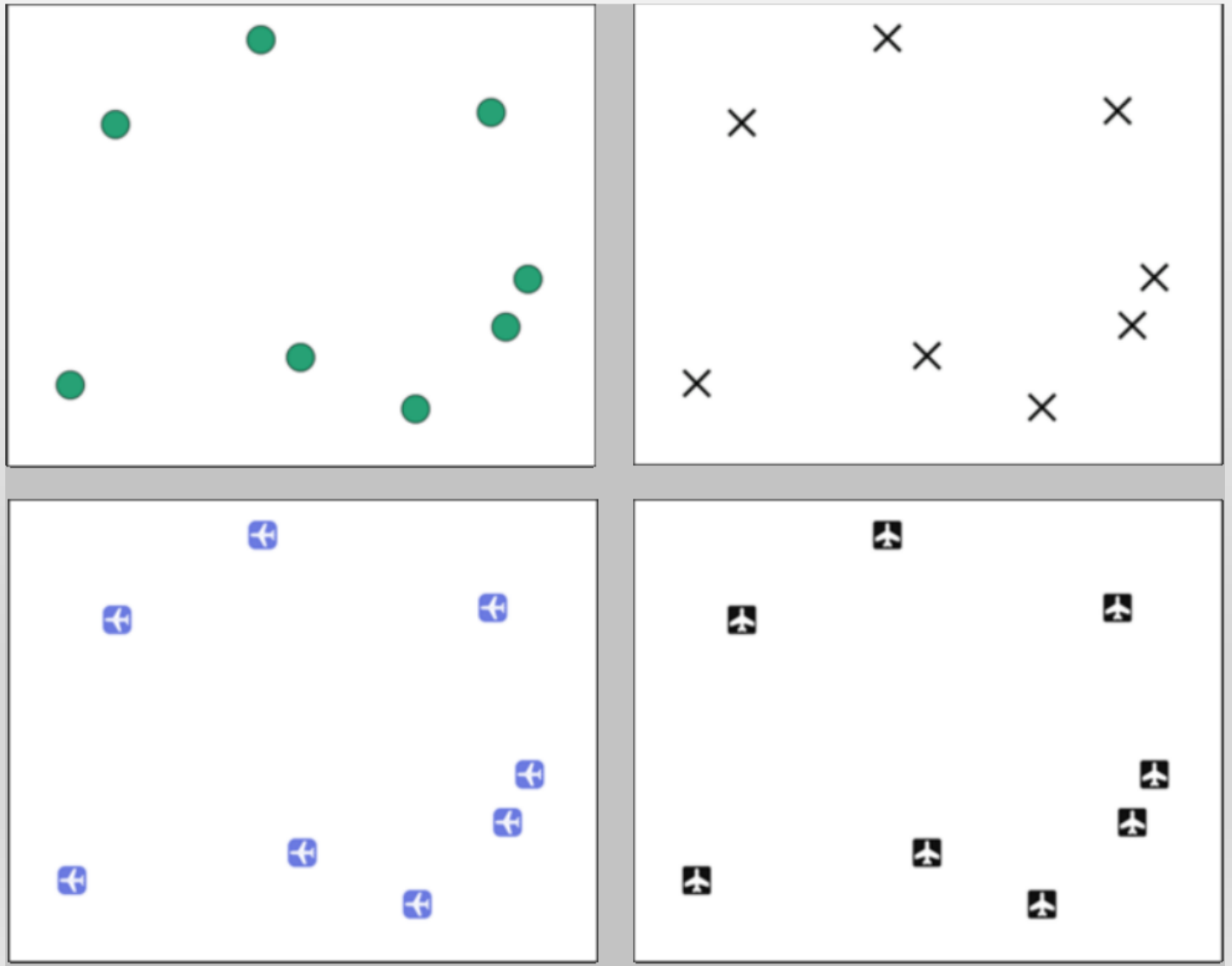
- Defines discrete objects
- Three Basic types of vector data
  - ✓Point
  - ✓Line
  - ✓Polygon

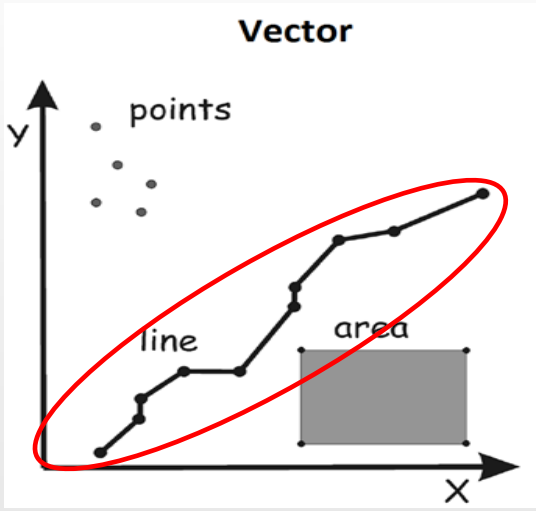
  } Simple features
- Composed of coordinates and attributes



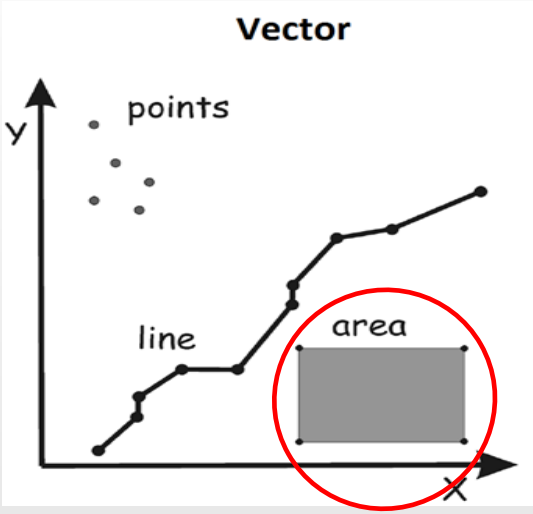
- **Point**
- Uses a single coordinate pair to define location
- Considered to have no dimension (they may have actual real-world dimensions, but for the purposes of a GIS, no dimension is assumed)
- Attribute information is attached to the point
- Examples: accident location, wells

- Different ways to represent airports



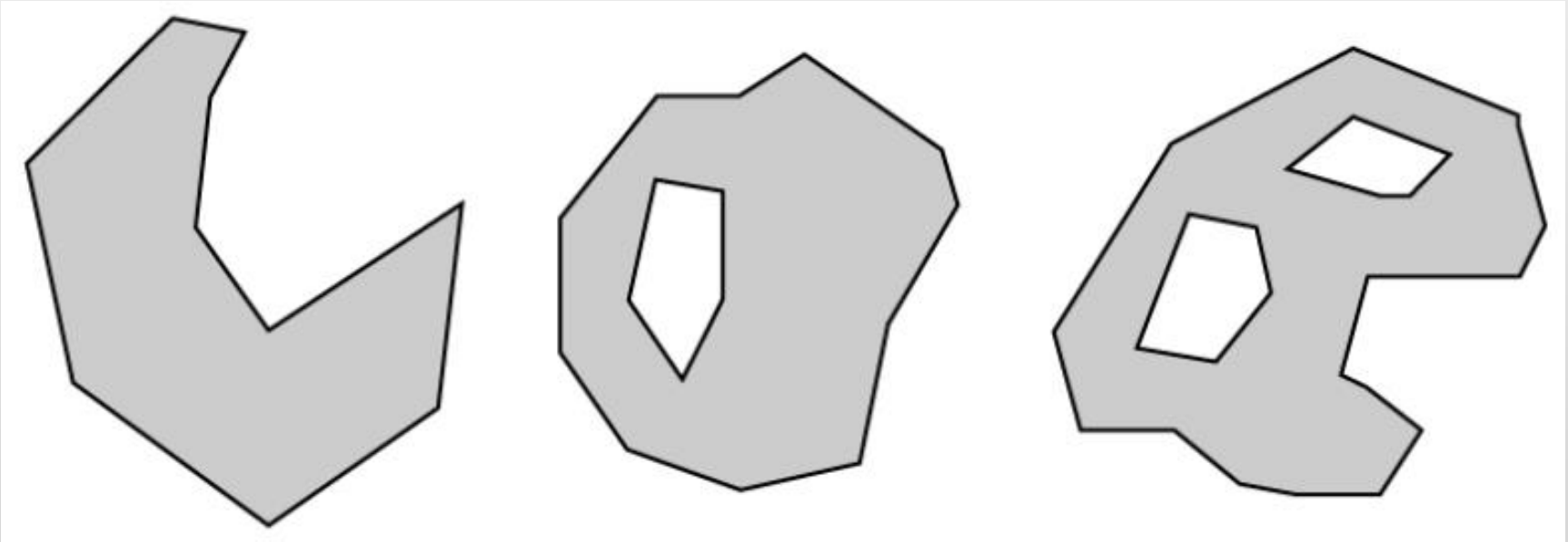


- **Line**
- Uses an ordered set of coordinates to define location
- Each line (and curve) is made up of multiple line segments
- Occasionally, curved lines are represented mathematically
- Starting point of a line is ***node***
- Intermediate point of a line is a ***vertex***
- Attributes may be attached to whole line , or node, or vertex
- Examples: road, streams



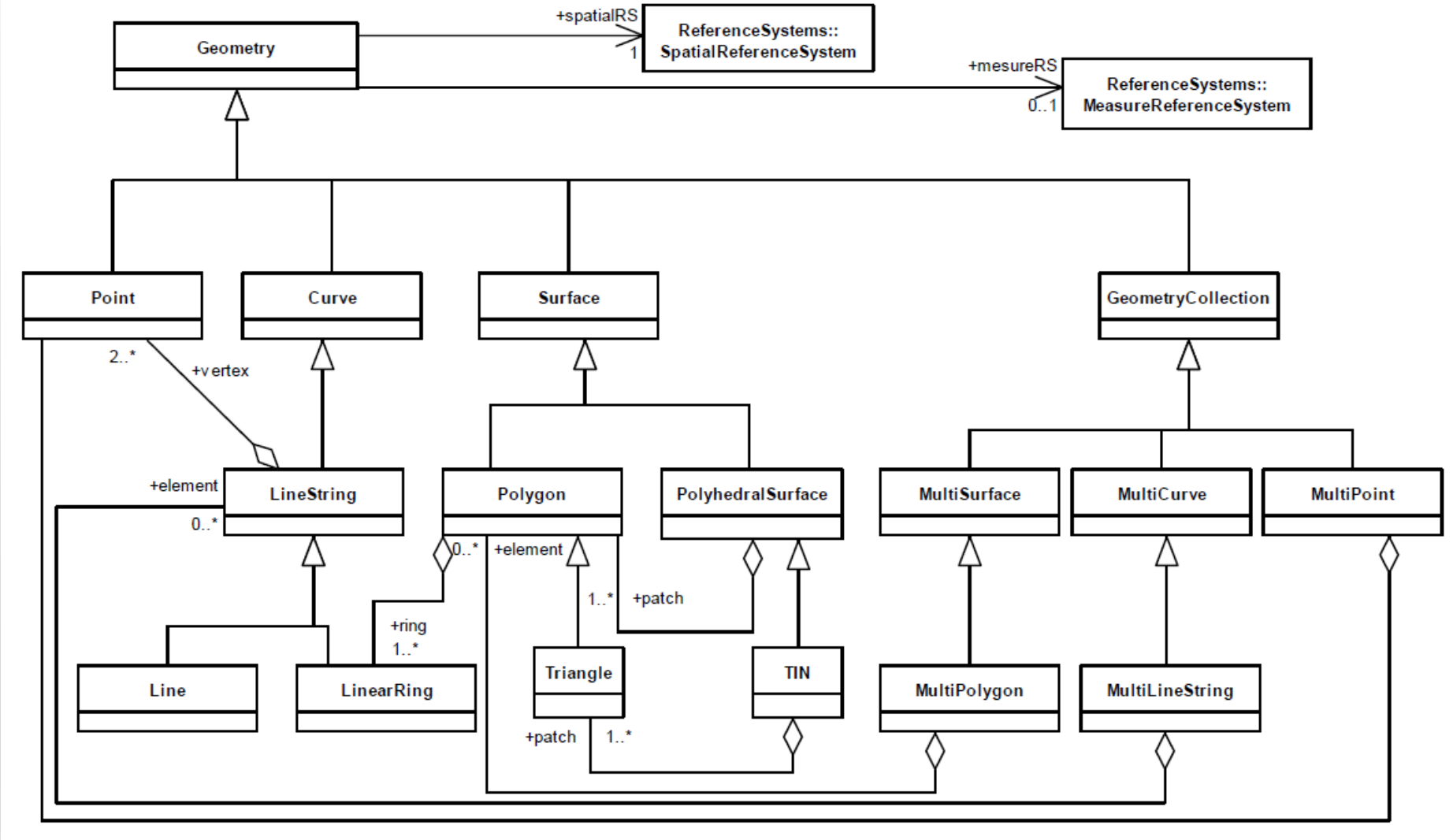
- **<u>Polygon</u>**
- Formed by a set of connected lines
- Polygons must close. The start and end point must have the same coordinate, or the polygon must close to an adjacent feature
- Polygons have an interior region
- Attribute information is attached to the polygon
- Examples: lake, city

- Examples of polygons



For more detailed description see: [OpenGIS® Implementation Standard for Geographic information - Simple feature Access]

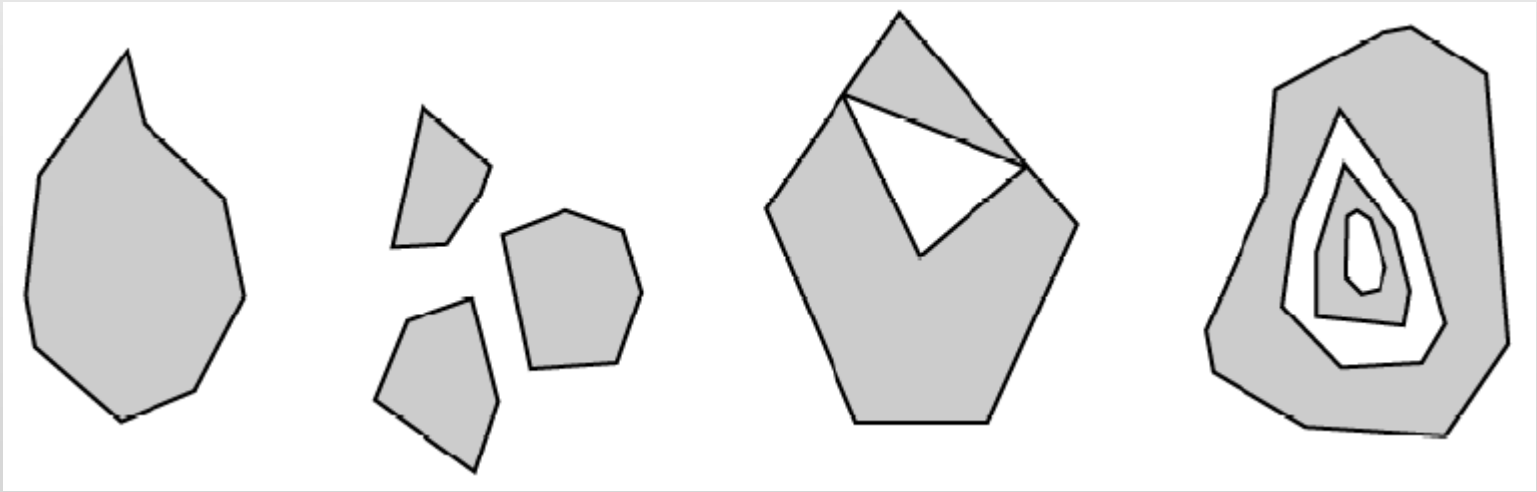
IN ADDITION TO POINT, LINE AND POLYGON…



- Geometry class hierarchy according to OpenGIS Simple Features Access (ISO 19125-1)

Source: http://portal.opengeospatial.org/files/?artifact_id=25355

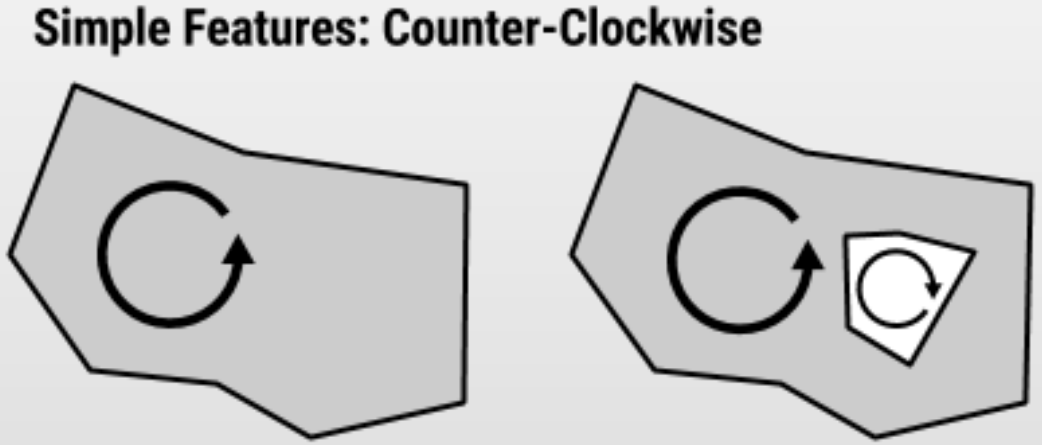
- **Multipolygon**
- A MultiPolygon is a MultiSurface whose elements are Polygons.



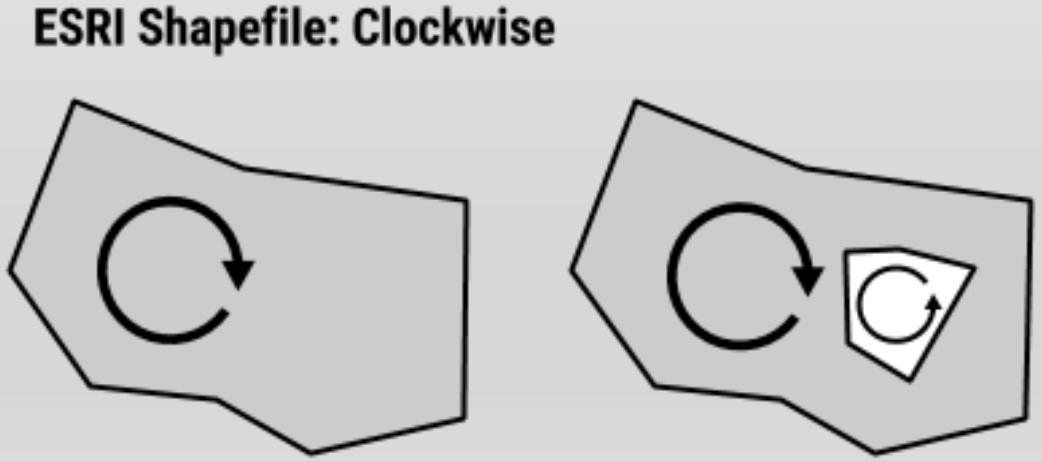
For more detailed description see: OpenGIS® Implementation Standard for Geographic information - Simple feature Access

Polygon boundary directions are needed to prevent ambiguities for geographic coordinate systems that cover a finite surface

Simple Feature Access (ISO 19125-1) also used in WKT/GML/KML and various SQL implementations:
exterior rings: counter-clockwise
interior rings (holes): clockwise direction.

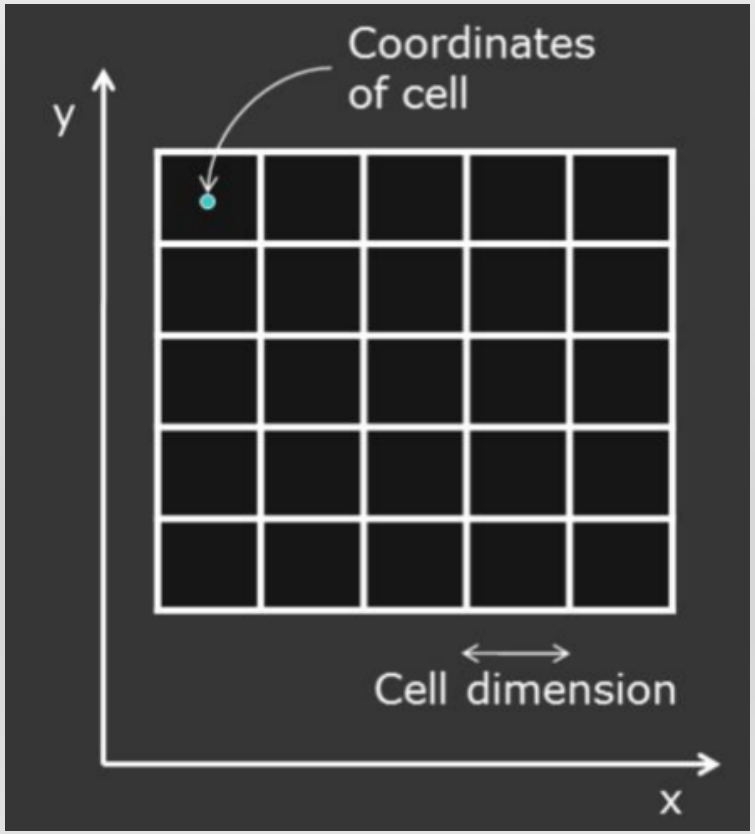


ESRI Shapefiles/ SHP:
exterior rings: clockwise
interior rings: counter-clockwise



RASTER DATA MODEL

- Represents usually continuous phenomena (temperature, elevation)
- Regular set of cells in a grid pattern
- Real-world objects are represented by value in the grid cell
- The cell size is the **resolution**
- There is a trade-off between resolution and raster file size
- The cell coordinate is the center point of the cell
- The coordinate applies to the entire cell area
- Each raster cell represents a given area and the value assigned applies to the entire cell
- The raster cell value represents the average, central, most common, or only value covered by the cell



- Satellite image of Estonia

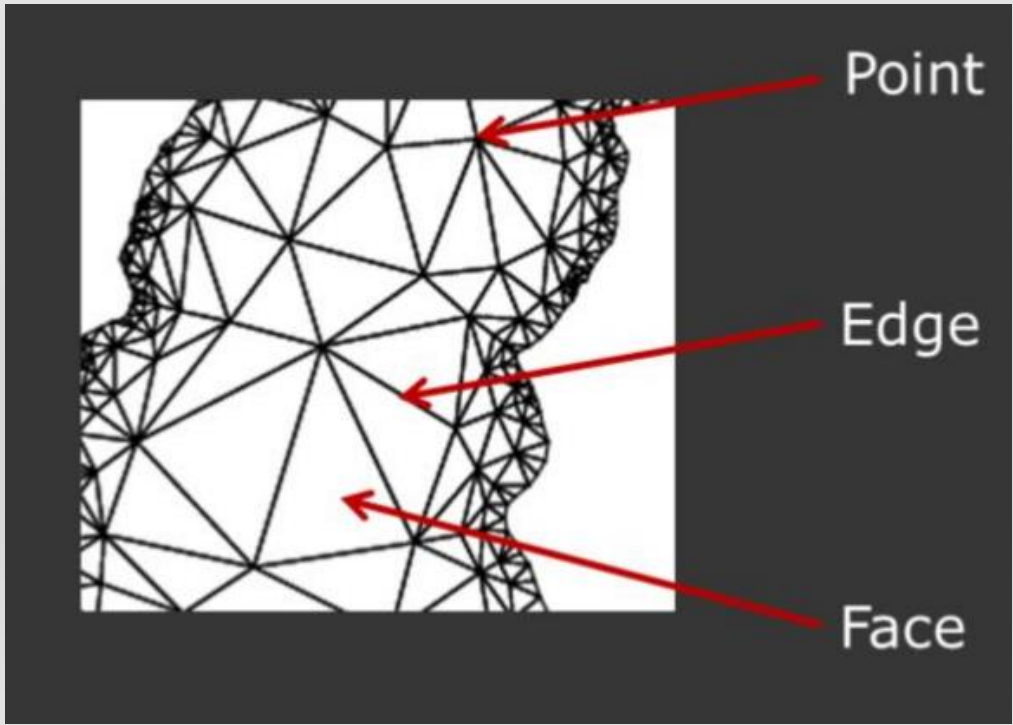
Raster vs Vector

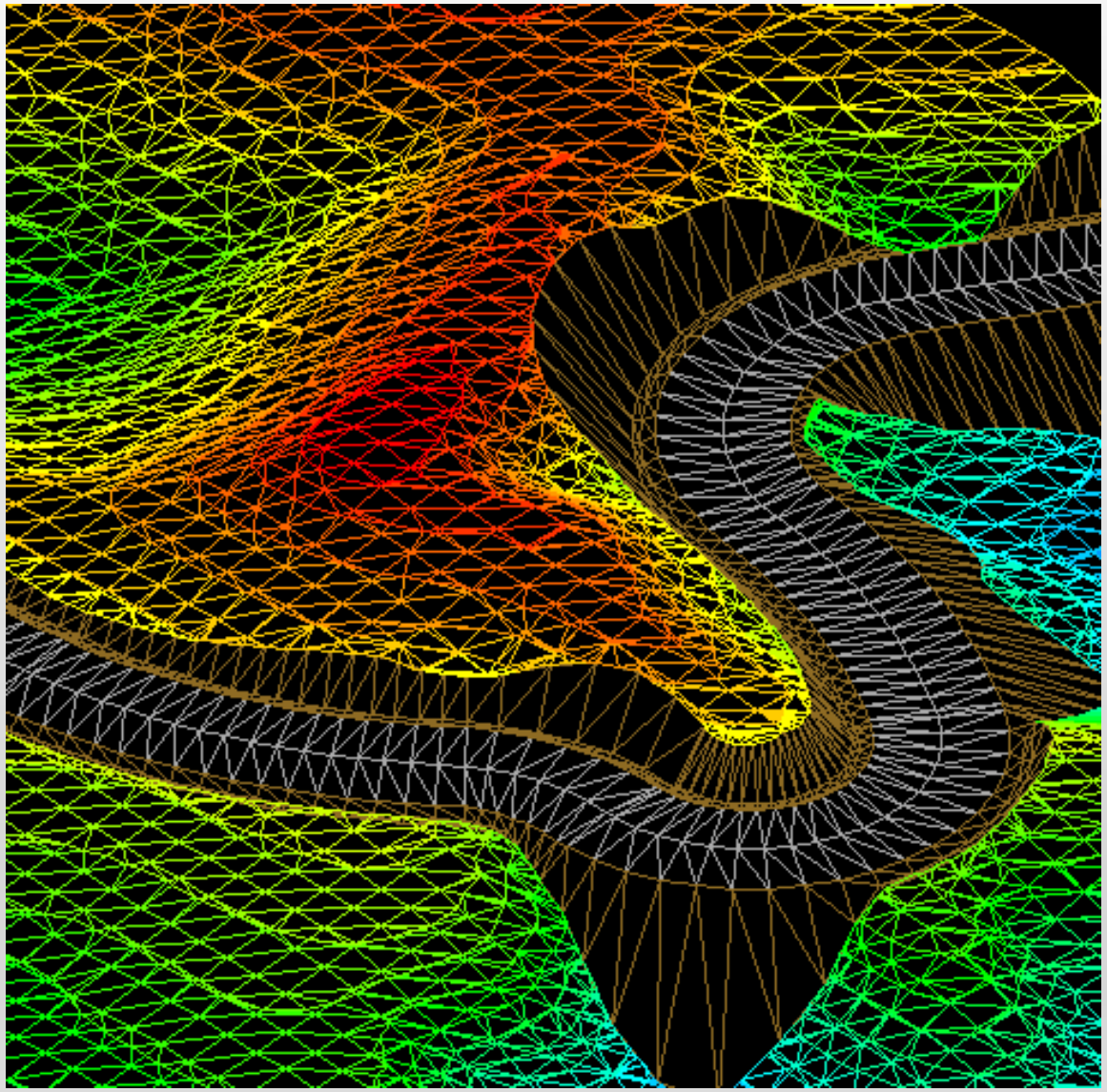
Raster	Vector
Good for frequent changes	Compact data storage
Simple data model	Great for network and linear features
Easy overlays, spatial analysis and modelling	Database management, query, reporting
Best for digital imagery	Can contain topology

Triangulated Irregular Networks (TIN)

- A Network of triangles connected together to create a 3D surface (triangles do not cross)
- More complex than rasters and more efficient space-wise
- Easily accommodates differing sample density
- TIN preserves each measurement point

Anatomy of a TIN





Source: http://www.surfacemodeling.com

METADATA

- **Metadata – „**data about data“
- **Geospatial metadata** is a type of metadata that is applicable to objects that have an explicit or implicit geographic extent, i.e. are associated with some position on the surface of the globe.
- **Purpose of metadata:** support discovery of data, and automated discovery, ingestion, processing and analysis
- **Metadata format:** usually xml
- **Metadata standards:** specifications for formatting and populating your metadata

More about metadata in the later lecture & lab session on Metadata

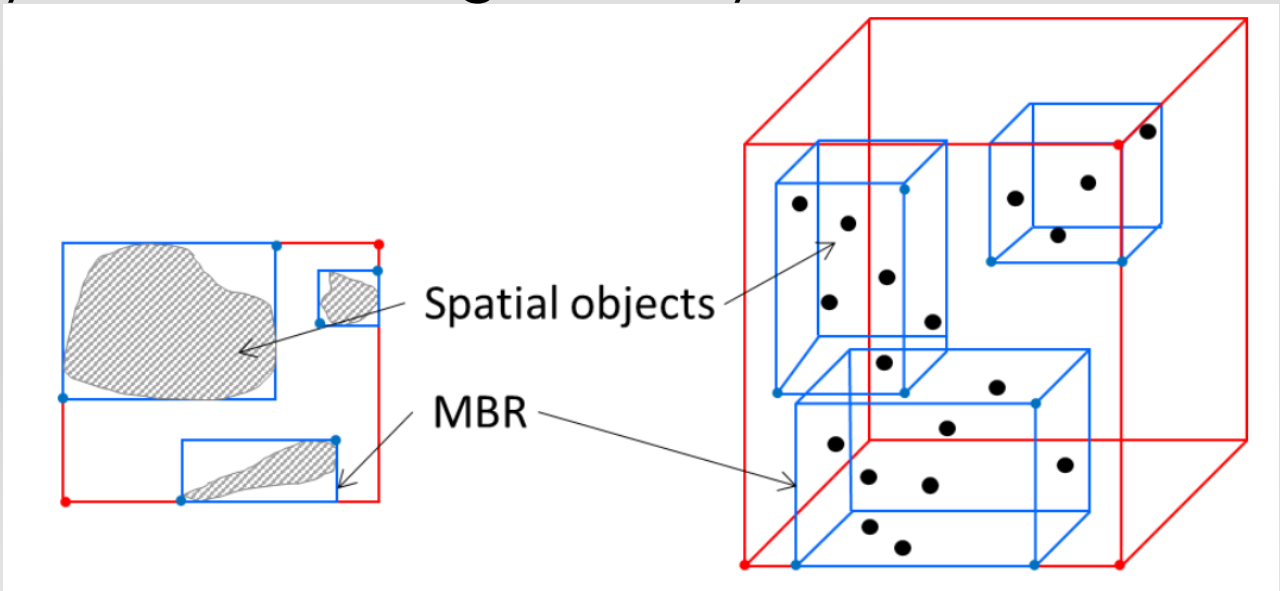
SPATIAL INDEXING

- A spatial index is a data structure that allows for accessing a spatial object efficiently.
- It is a common technique used by spatial databases.
- Without indexing, any search for a feature would require a "sequential scan" of every record in the database, resulting in much longer processing time.

Source: Zhang, X and Du, Z. (2017)

SPATIAL INDEXING
BOUNDING BOX

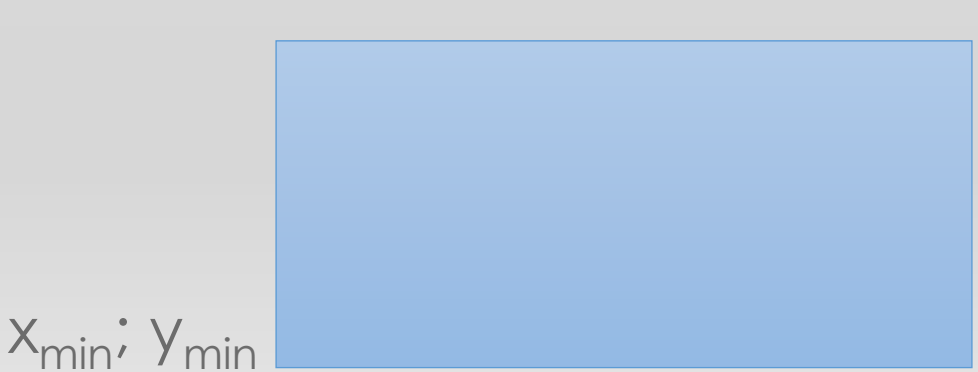
- In order to reduce the cost of calculating the complex shape of spatial object during the search traversal, we use approximations of the complex object geometries.
- The most commonly used approximation is the **minimum bounding rectangle**, or MBR, which is also called **minimum bounding box**, or MBB. An MBR is a single rectangle that minimally encloses the geometry.



Source: Zhang, X and Du, Z. (2017)

BOUNDING BOX

- In a 2D plane, an MBR is defined by four coordinates, (x_{min} y_{min}) and ($x_{max}y_{max}$). These coordinates represent the following:
  - x_{min} is the x-coordinate of the lower-left corner of the bounding box.
  - y_{min} is the y-coordinate of the lower-left corner of the bounding box.
  - x_{max} is the x-coordinate of the upper-right corner of the bounding box.
  - y_{max} is the y-coordinate of the upper-right corner of the bounding box.



Source: Zhang, X and Du, Z. (2017)

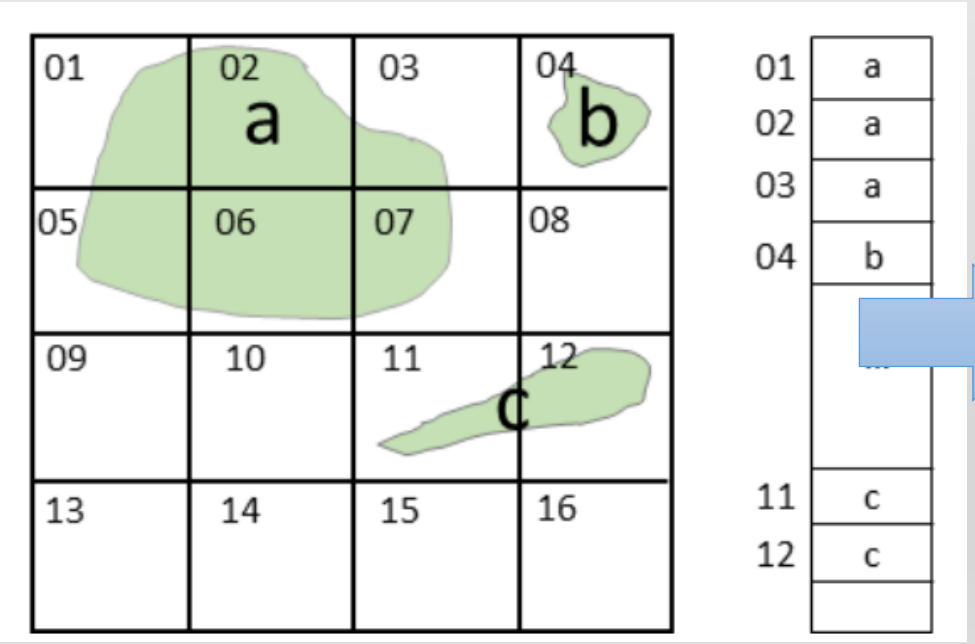
SPATIAL INDEXING
DIFFERENT METHODS/TYPES OF SPATIAL INDEXING

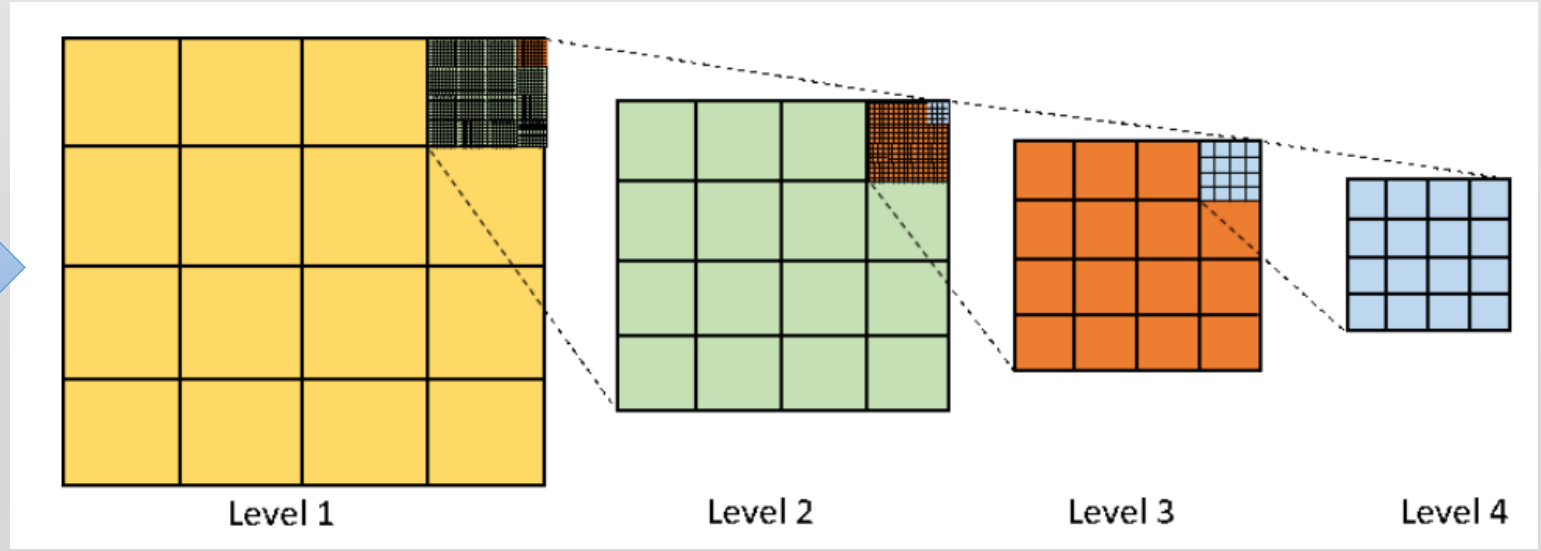
- Different data sources use different data structures and access methods. Two well-known spatial indices:
  - **Space-driven structures.** These data structures are based on partitioning of the embedding 2D space into cells (or grids), mapping MBRs to the cells according to some spatial relationship (overlap or intersect). Microsoft SQL Server, ESRI geodatabase use these methods.
  - **Data-driven structures.** These data structures are directly organized by partition of the collection of spatial objects. Data Objects are grouped using MBRs adapting to their distribution in the embedding space. PostGIS, MySQL, QGIS gpkg use these data structures.

Source: Zhang, X and Du, Z. (2017)

SPACE-DRIVEN STRUCTURES

1. Fixed grid index is an n×n array of equal-size cells. Each one is associated with a list of spatial objects which intersect or overlap with the cell.

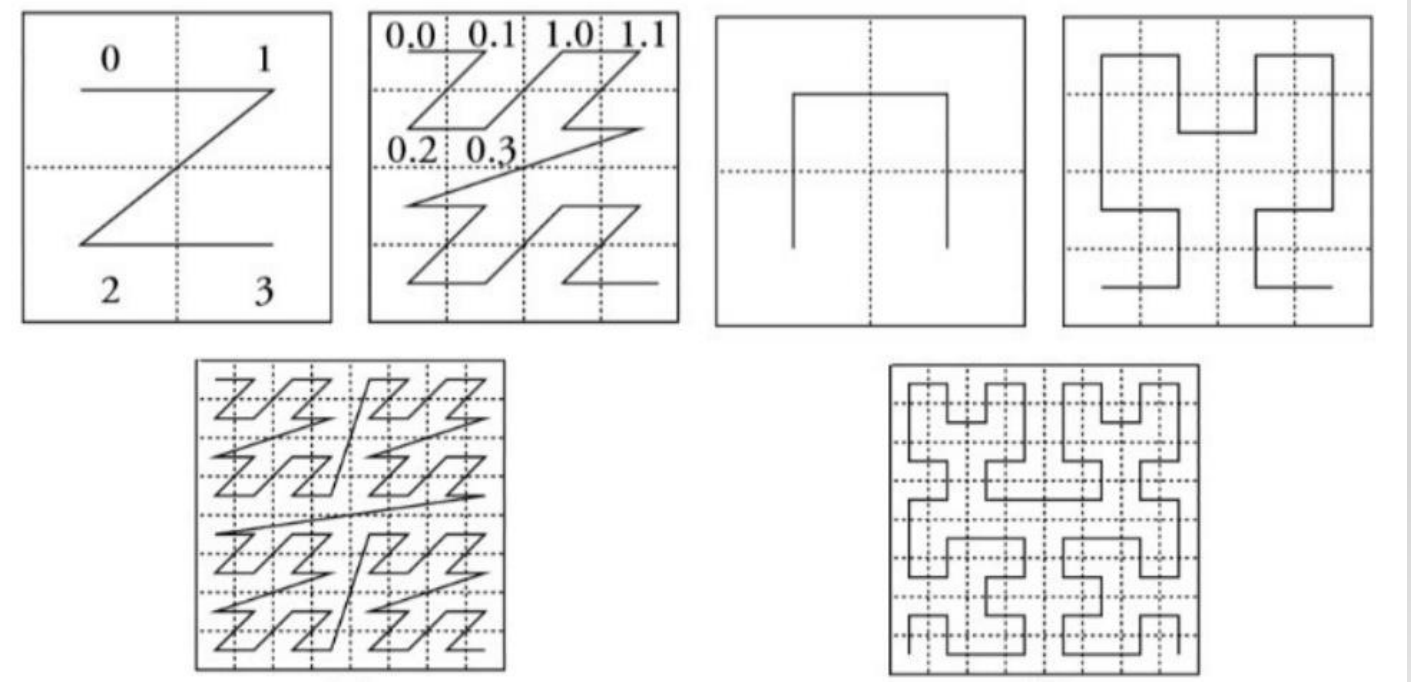




Source: Zhang, X and Du, Z. (2017)

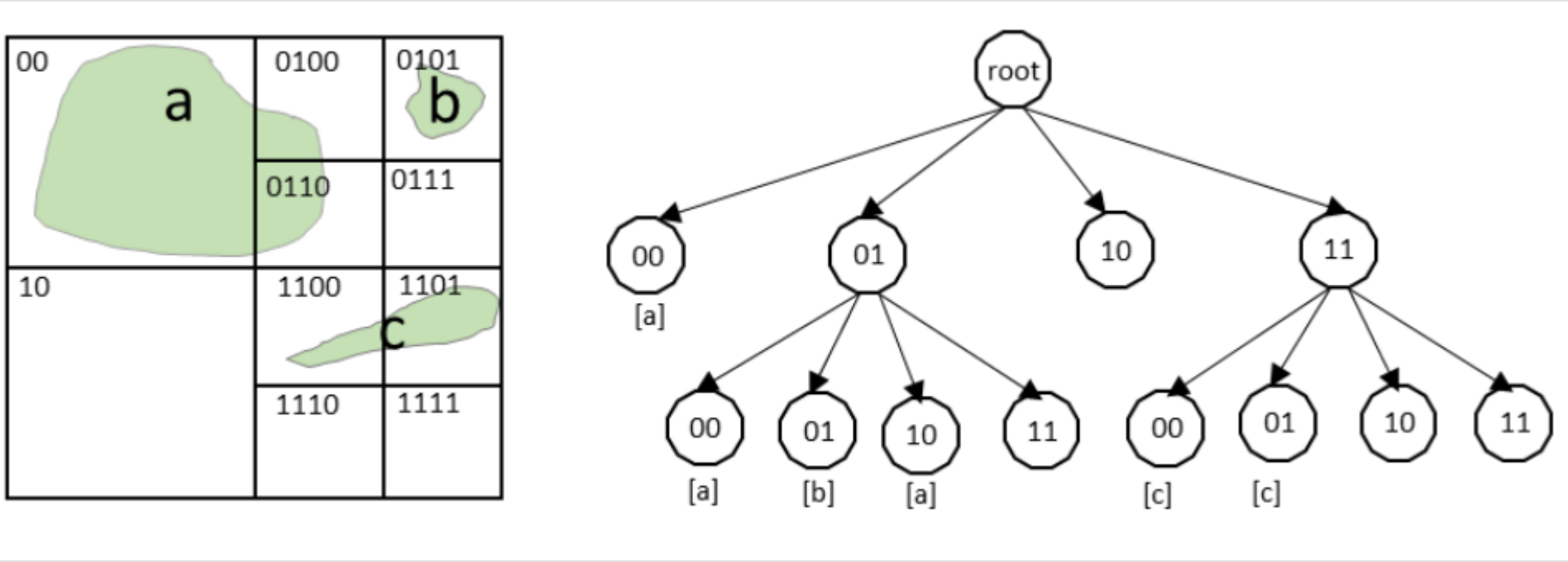
- These grid hierarchy cells (previous slide) are numbered in a linear fashion called space-filling curves. They are useful because it partially preserves proximity, that is, two cells close in 2D plane are likely to be close in the sequential order.
- There are different spatial filling curves:

z-order *Hilbert curve*



Source: Zhang, X and Du, Z. (2017)

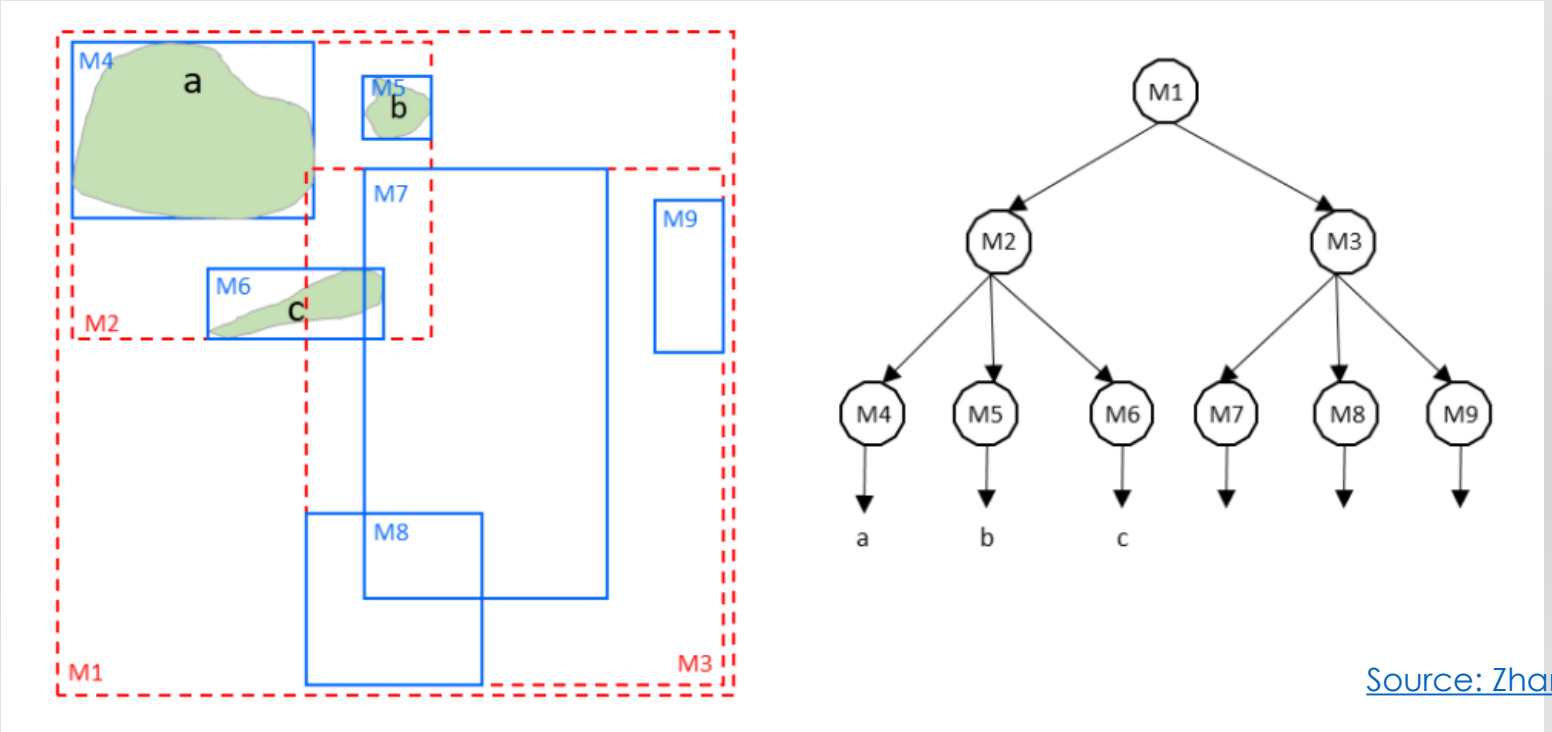
2. Quadtree is a very popular spatial indexing technique. It is a specialized form of grid in which the resolution of the grid is varied according to the density of the spatial objects to be fitted.



Source: Zhang, X and Du, Z. (2017)

DATA-DRIVEN STRUCTURES

R-tree index consists of a hierarchical index on the MBRs of the geometries in the layer of geometries. This hierarchical structure is based on the heuristic optimization of the area of MBRs in each node in order to improve the access efficiency.



Source: Zhang, X and Du, Z. (2017)

- A **space-driven spatial** index has the advantage that the structure of the index can be created first, and data will then be added gradually without requiring any change to the index structure; indeed, if a common grid is used by disparate data collecting and indexing activities, such indices can easily be merged from a variety of sources.
- **Data-driven structures** such as R-trees can be more efficient for data storage and faster in search execution time, but they are generally tied to the internal structure of a given data storage system.

Source: Zhang, X and Du, Z. (2017)

DATA COMPRESSION

- Data compression principles:
  - ✓Is the substitution of frequently occurring data items, or symbols, with short codes that require fewer bits of storage than the original symbol.
  - ✓Saves space, but requires time to save and extract.
  - ✓Success varies with type of data.
  - ✓Works best on data with low spatial variability and limited possible values.
  - ✓Works poorly with high spatial variability data or continuous surfaces.
  - ✓Exploits inherent redundancy and irrelevancy by transforming a data file into a smaller one.

Source: Dolci et al. Structures for Data Compression

DATA COMPRESSION: LOSSLESS AND LOSSY

- **Lossless compression** algorithm eliminates only redundant information, so that one can recover the data exactly upon decompression of the file. **Lossless data compression is compression without any loss of data quality.** The decompressed file is an exact replica of the original one. Lossless compression is used when it is important that the original and the decompressed data are identical. Some image file formats, notably PNG, use only lossless compression, while those like TIFF may use either lossless or lossy methods.
- **Examples of lossless methods** are: run-length coding, Huffman coding, Lempel-Ziv-Welsh (LZW) method

Source: Dolci et al. Structures for Data Compression

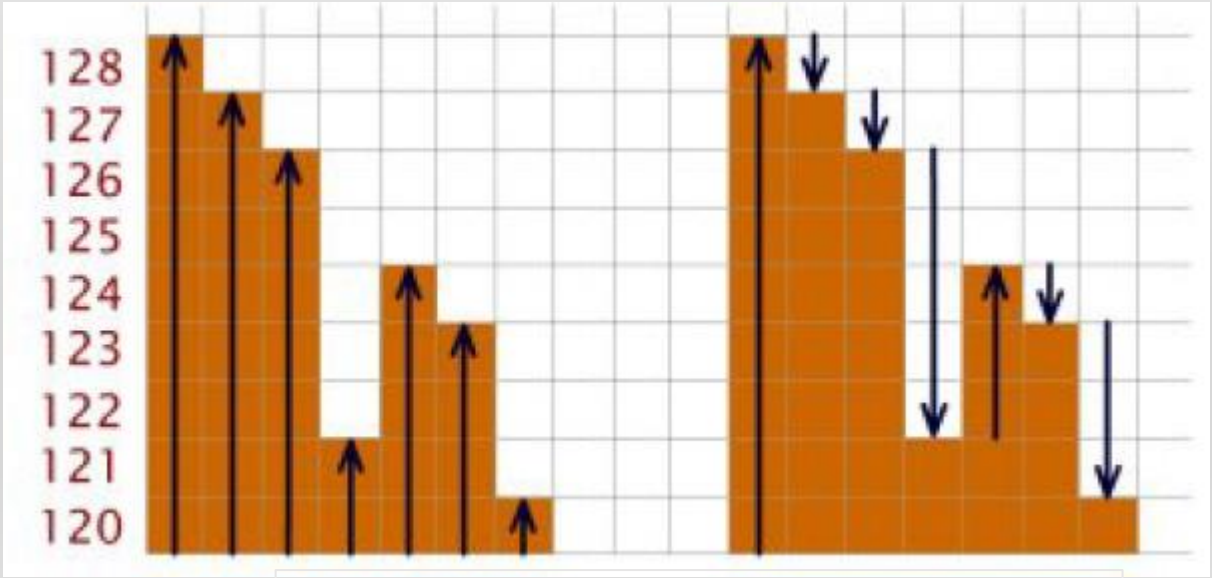
Numerical example of lossless method:

An example:

128, 127, 126, 121, 124, 123, 120

Can be re-written in shorter numbers requiring less bits like:

128, -1, -1, -5, +3, -1, -3



Lossless compression

Source: Dolci et al. Structures for Data Compression

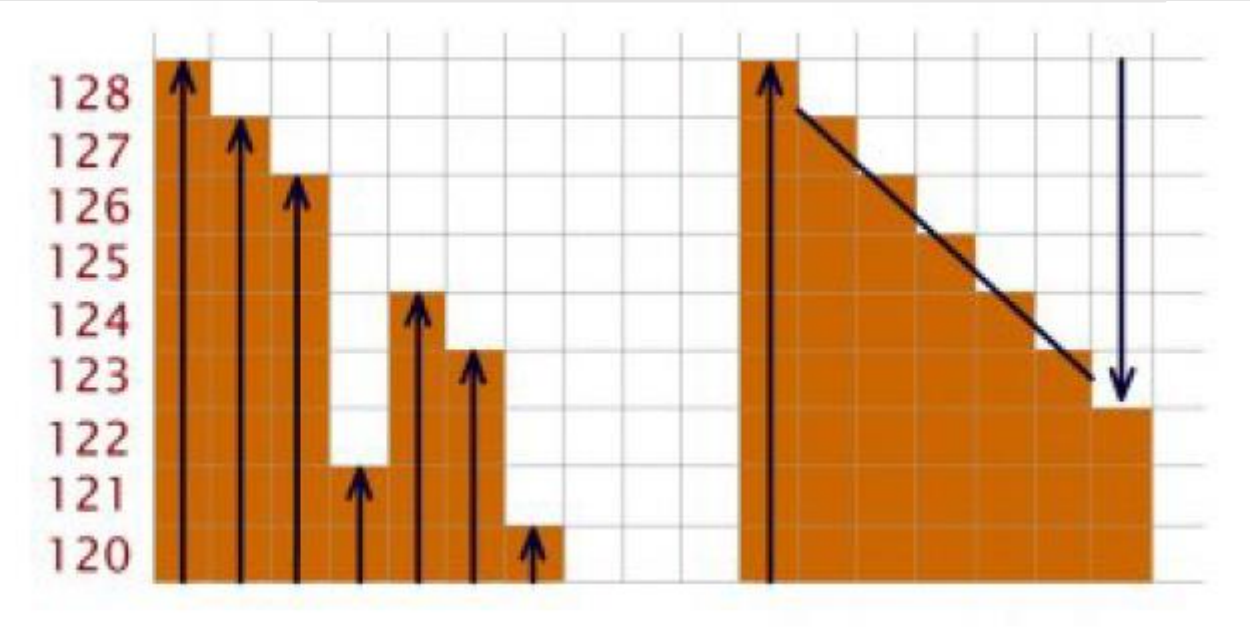
- **Lossy compression** method is one where compressing data and then decompressing it retrieves data that may well be different from the original, but is "close enough" to be useful in some way. The algorithm eliminates irrelevant information as well, and permits only an approximate reconstruction of the original file. Lossy compression is also done by re-writing the data in a more space efficient way, but more than that: less important details of the image are manipulated or even removed so that higher compression rates are achieved.
- Lossy compression is dangerously attractive because it can provide compression ratios of 100:1 to 200:1, depending on the type of information being compressed. But the cost is loss of data.
- Examples of LOSSY METHODS are: PCM, JPEG, MPEG

Source: Dolci et al. Structures for Data Compression

Numerical example of lossy method:

The previous sequence of numbers
128, 127, 126, 121, 124, 123, 120
can be re-written like:
128 - 6
Result after decompression:
128, 127, 126, 125, 124, 123, 122

Lossy compression



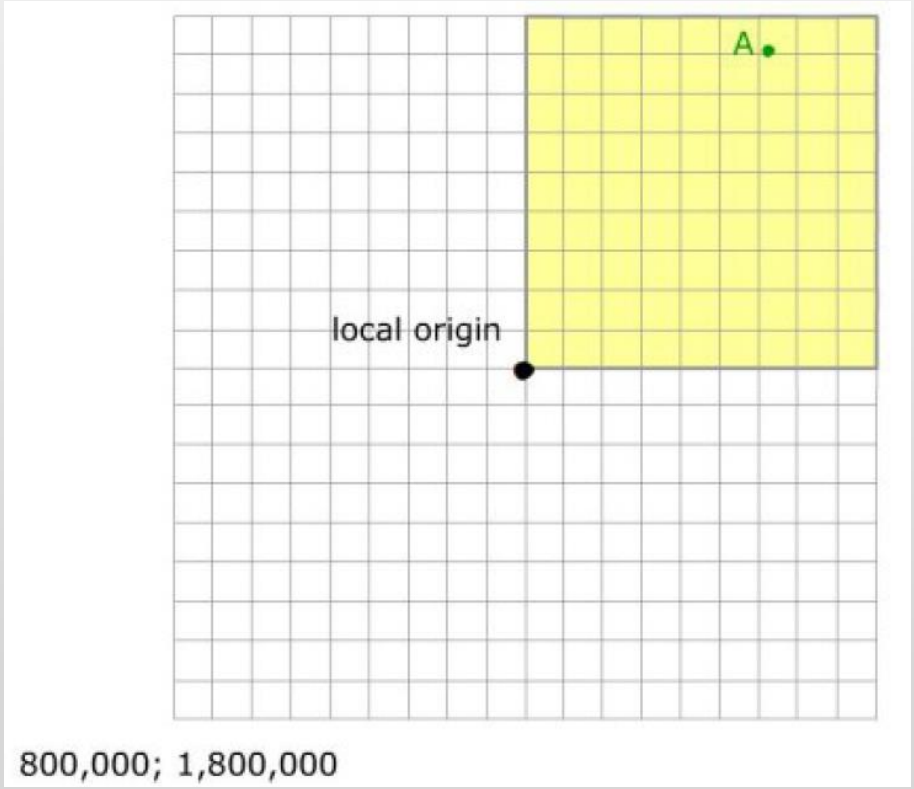
Source: Dolci et al. Structures for Data Compression

VECTOR DATA COMPRESSION

For example, Point A (897 345.32; 1 898 765.98)

To save storage space, it is possible to define a **LOCAL ORIGIN** and store all the coordinates relative to the new origin. These new coordinates will be smaller numbers than the original ones and can be stored in a smaller amount of computer storage space.

The coordinate offsets will preserve the **ABSOLUTE COORDINATES,** which will be used to restore the information when we use the data.

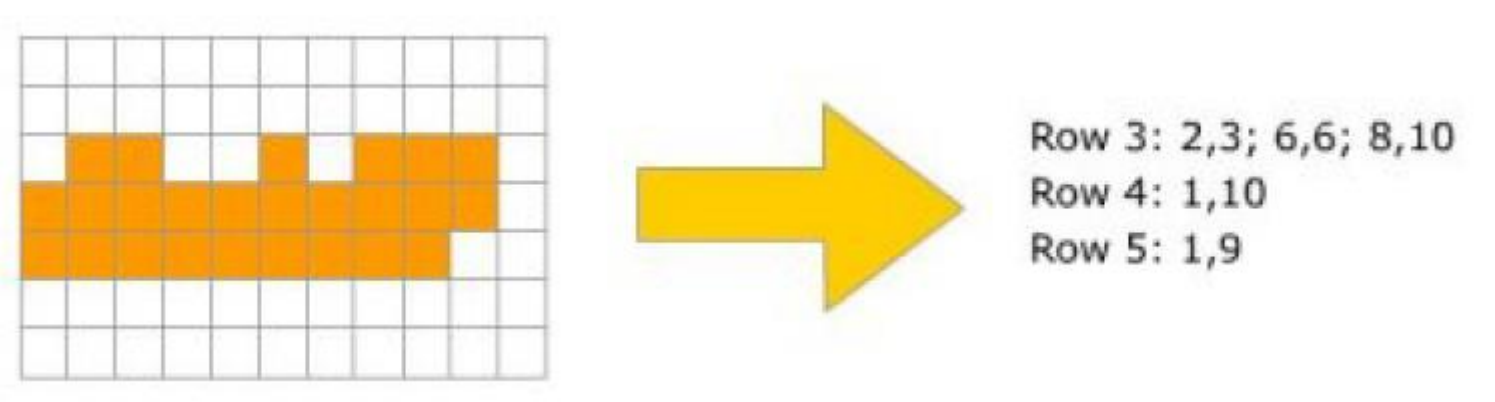


Source: Dolci et al. Structures for Data Compression

RASTER DATA COMPRESSION I

Runlength coding (lossless)

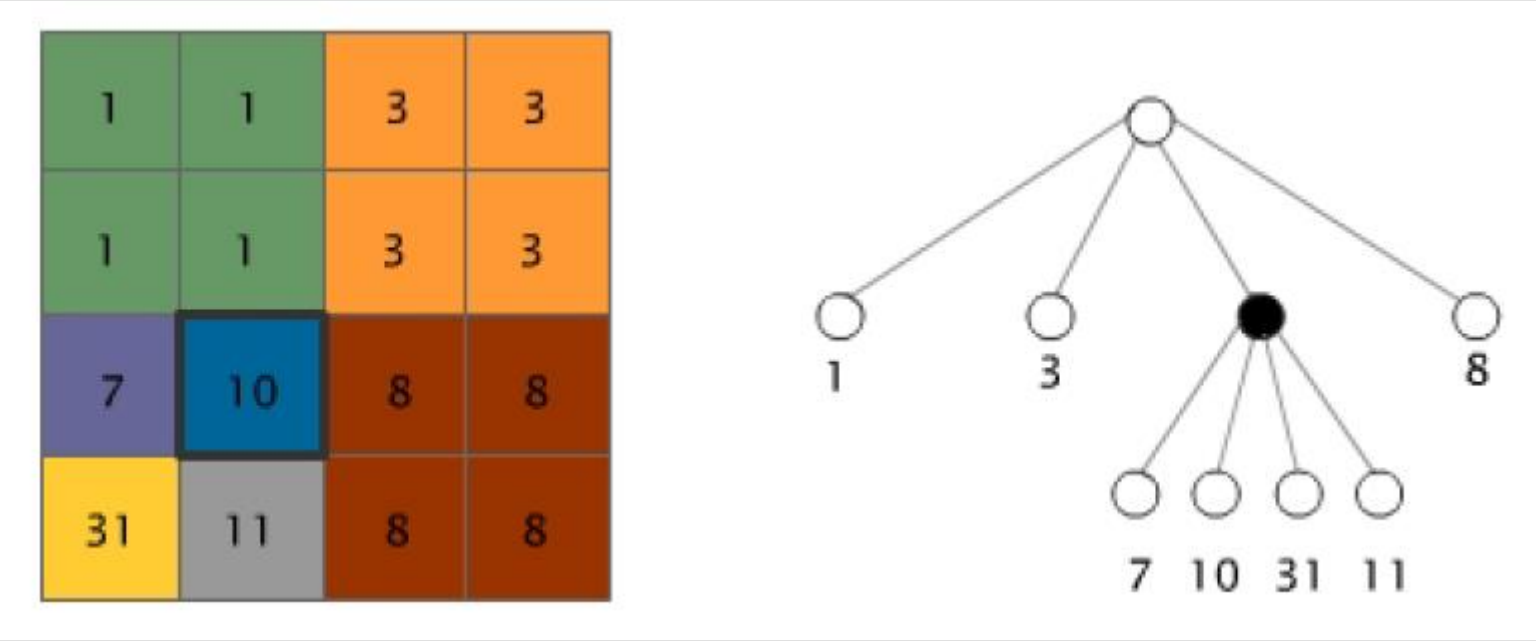
- Geographical data tends to be "spatially autocorrelated", meaning that objects which are close to each other tend to have similar attributes. Therefore, instead of repeating pixel values, we can code the raster as pairs of numbers - (run length, value).
- The runlength coding is a widely used compression technique for raster data. The primary data elements are pairs of values or tuples, consisting of a pixel value and a repetition count which specifies the number of pixels in the run. Data are built by reading successively row by row through the raster, creating a new tuple every time the pixel value changes or the end of the row is reached.



Source: Dolci et al. Structures for Data Compression

RASTER DATA COMPRESSION II

- **The quadtree compression** technique is the most common compression method applied to raster data.
- Quadtree coding stores the information by subdividing a square region into quadrants, each of which may be further subdivided in squares until the contents of the cells have the same values.



Source: Dolci et al. Structures for Data Compression

RASTER DATA COMPRESSION III

- LZ77 compression (also known as LZ or LZW) method is relatively simple: when you find a match (a data value that has already been seen in the input file) instead of writing the actual value, the position and length (number of bytes) of the value is written to the output (the length and offset - where it is and how long it is).
- Method applicable to all the raster data types.

DATA ORGANISATION

- Keeping your data organised is a huge factor in the project management. **The speed at which a project is completed can be increased significantly by merely keeping data organised.** When it is easy to find data, the process is sped up for both you and other involved parties. Poorly organised data not only slows down projects but also frustrates everyone at the same time. Additionally, as a project is passed around, it makes it easier for the next person to pick up where the previous person left off it is well organised.

- There are many ways to organise your data for a project. At a minimum, you should create **specific folders to sort your data**, for example, placing all your shapefiles for a project in a single-parent directory, or even separating them by theme. You can also copy items to keep the original raw data without any edits in case you need to revert back to the original data set.

- **Items that are no longer used such as intermediate output from tool runs should be deleted instead of kept.** If you have a large amount of data and tools, you may also consider using a database to organise the information. **And lastly, you should rename items to make recognition easier.**

ASSIGN DESCRIPTIVE FILE NAMES

smart_tartu_2017_precip.tif

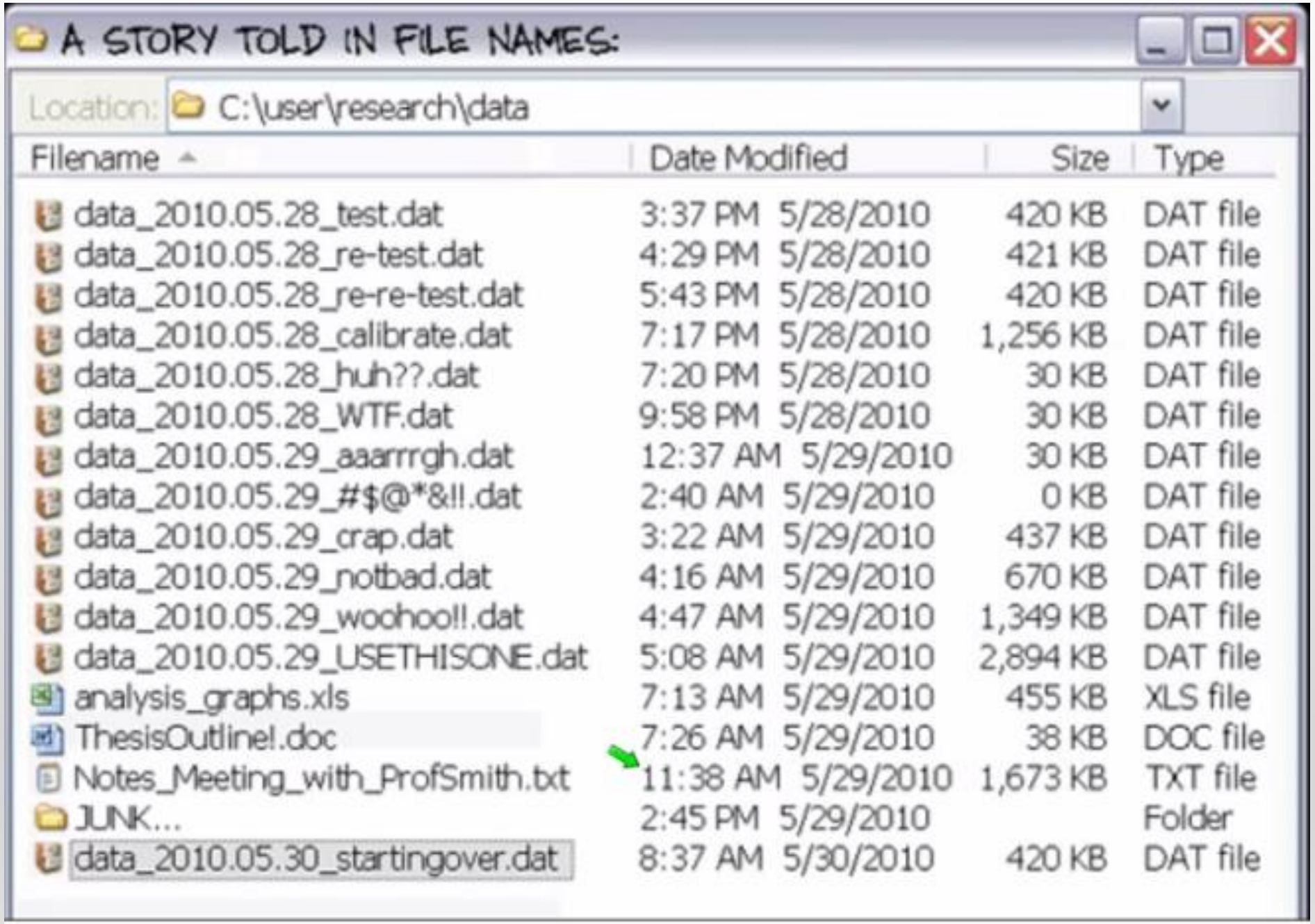
File names should:

- Be unique
- Reflect contents
- Use ASCII characters only
- **Avoid spaces and special characters**
- Try to follow your communities' rules

Bad examples:
My data.xls
2017 data.csv
Final version.csv
Etc.

FOR EXAMPLE, CLIMATE&FORECAST (CF) STANDARD NAMES

- canopy_water_amount
- change_over_time_in_surface_snow_amount
- convective_precipitation_amount
- convective_precipitation_flux
- convective_precipitation_rate
- convective_rainfall_amount
- convective_rainfall_flux
- convective_rainfall_rate
- convective_snowfall_amount
- convective_snowfall_flux
- correction_for_model_negative_specific_humidity
- downward_heat_flux_at_ground_level_in_snow
- eastward_atmosphere_water_transport_across_unit_distance
- eastward_atmosphere_water_vapor_transport_across_unit_distance
- effective_radius_of_convective_cloud_rain_particle
- effective_radius_of_convective_cloud_snow_particle
- effective_radius_of_stratiform_cloud_rain_particle
- effective_radius_of_stratiform_cloud_snow_particle
- heat_flux_into_sea_water_due_to_snow_thermodynamics
- humidity_mixing_ratio
- integral_of_product_of_eastward_wind_and_specific_humidity_wrt_height
- integral_of_product_of_northward_wind_and_specific_humidity_wrt_height
- land_ice_runoff_flux
- liquid_water_content_of_surface_snow
  alias: liquid_water_content_of_snow_layer



Courtesy of PhD Comics

PRESERVE PROCESSING INFORMATION

Keep raw data raw:

- Do not include transformations, interpolations, etc. in a raw file
- Make your raw data „read only“ to ensure no changes
- When processing data:
  - Write documentation and when using deskdop GIS-programs then it is useful to show processing type also in the file name
  - More efficient is to use programming laguage (R, Python etc) and version control (Git, Subversion etc)

PERFORM BASIC QUALITY CONTROL

- Check missing values or extremes (outliers) which might be also impossible
- Avoid using NaN and do not definitely use 0 (zero), use a Missing Value Code (-9999) instead
- Perform and review statistical summaries to find errors in data

WELL ORGANISED DATA:

- Enables to work more efficiently
- Can be shared easily by collaborators
- Can be potentially be re-used

Include data management in your workflow!

BENEFITS FROM MAKING YOUR DATA EASILY ACCESSIBLE AND READABLE

- Promote sharing and research
- Benefit from the information infrastructures that are provided by others (NASA, NOAA; Google etc.)
- Your data can be ingested into many existing Web services to provide on-demand data distribution to users.

KEY „TAKE-HOME MESSAGES“

- Provide geospatial, temporal, other information completely and accurately
- Choose good formats to organise the data content and make them self-descriptive
- Provide metadata in standard ways
- There are many benefiits of using well-organised data

LAB SESSION

- In the lab session, you will become familiar with different data models and data types.
- Lab session has three parts. A and B are compulsory; C is optional (advanced).
- The optional part (C) can earn you 2 extra points

QUIZZES

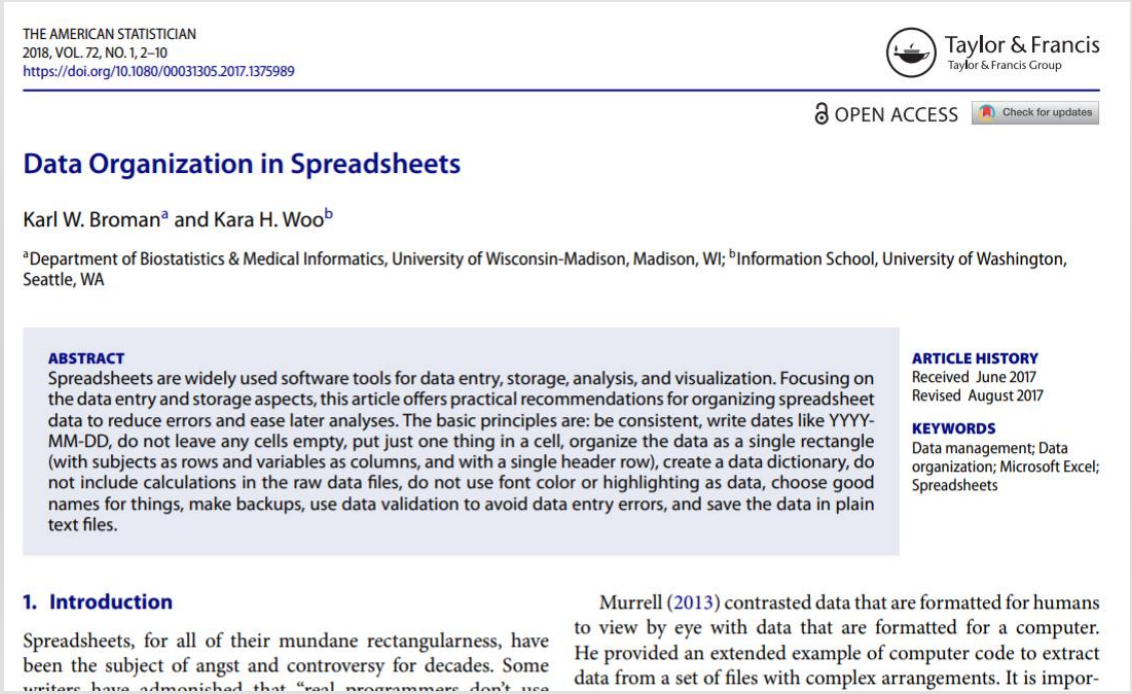
- While doing the tutorial, you need to occasionally answer quiz in Moodle which are meant as self-check and learning if you understood the topic correctly.
- You can try the quizzes as many times as you wish and use them also to prepare for the test, except for optional tasks' quizzes.
- Optional tasks quizzes have only 1 trial.

INDEPENDENT TASK 1

- Review one application area of GIS. The application review can be prepared from published literature, WWW, and other media.
- You will present the application review as oral presentation on next seminar and upload your presentation to Moodle in **pdf-format.**
- **You presentation time is max 5 mins and you can have max 5 slides!**
- Presentations and submission: Friday, 17. September 2020
- Presentations will be in two groups

ADDITIONAL READING IN MOODLE

- Chapter 1 and 2 from Bolstad (2008).
- Dolci et al. 2010. Stuctures of Data Compression
- Broman and Woo, 2018. Data Organization in Spreadsheets



THE AMERICAN STATISTICIAN
2018, VOL. 72, NO. 1, 2–10
https://doi.org/10.1080/00031305.2017.1375989

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

Data Organization in Spreadsheets

Karl W. Broman[a] and Kara H. Woo[b]

[a]Department of Biostatistics & Medical Informatics, University of Wisconsin-Madison, Madison, WI; [b]Information School, University of Washington, Seattle, WA

ABSTRACT

Spreadsheets are widely used software tools for data entry, storage, analysis, and visualization. Focusing on the data entry and storage aspects, this article offers practical recommendations for organizing spreadsheet data to reduce errors and ease later analyses. The basic principles are: be consistent, write dates like YYYY-MM-DD, do not leave any cells empty, put just one thing in a cell, organize the data as a single rectangle (with subjects as rows and variables as columns, and with a single header row), create a data dictionary, do not include calculations in the raw data files, do not use font color or highlighting as data, choose good names for things, make backups, use data validation to avoid data entry errors, and save the data in plain text files.

ARTICLE HISTORY
Received June 2017
Revised August 2017

KEYWORDS
Data management; Data organization; Microsoft Excel; Spreadsheets

1. Introduction

Spreadsheets, for all of their mundane rectangularness, have been the subject of angst and controversy for decades. Some

Murrell (2013) contrasted data that are formatted for humans to view by eye with data that are formatted for a computer. He provided an extended example of computer code to extract data from a set of files with complex arrangements. It is impor-

REFERENCES

- Bolstad, P., 2008. GIS Fundamentals: A First Text on Geographic Information Systems, 3rd edition, Eider Press, 620 pp.
- Zhang, X and Du, Z. (2017). Spatial Indexing. The Geographic Information Science & Technology Body of Knowledge (4th Quarter 2017 Edition), John P. Wilson (ed). DOI: 10.22224/gistbok/2017.4.12
- Dolci et al. 2010. Stuctures of Data Compression
- FOSS4G GeoAcademy Curriculum
- Humboldt State University Geospatial online materials
- Stevens, S. S. 1946. On the Theory of Scales of Measurement. Science. 103 (2684): 677–680.