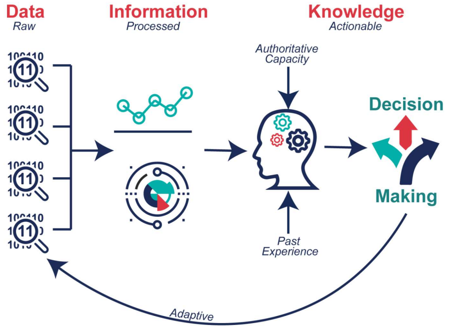
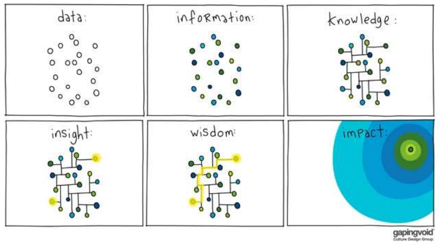
September 23: Getting to know Data with Python

Friday, August 20, 2021 6:54 PM

Data are at the heart of data science and, in fact, all scientific endeavors. Every discipline has distinct ways to collect and process data for analysis so that the data are appropriate to address its research questions. In other words, what we want to know determine what data we will collect.

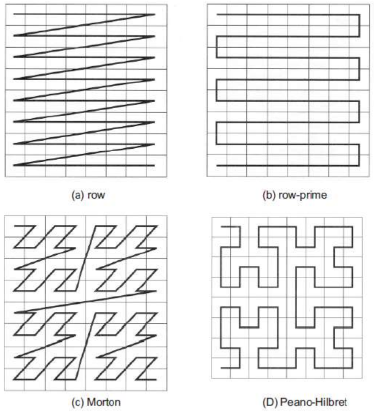
Data are observations or facts, representing what is known. Data science systematically study the structure and behavior of data in order to quantifiably understand the past and the present and predict the future possibilities. In data science, the data determine what the past and the present you can know.





Data have many dimensions:

1. Data type:
   a. Categorical vs. numerical
   b. Nominal, ordinal, interval, and ratio
   c. Text, integer, real (float)
2. Data format
   a. dd-mm-yyyy or MM/DD/YY
   b. Precision: %.2f (two decimal digits)
   c. Latitude and longitude in DMS or decimal degree
   d. Capitalization: all caps, all lowercase, or capitalize the first letter
3. Data structure
   a. A value
   b. Set ---> (any items no order)
   c. List ---> [any items in order]
   d. Array ---> [numbers only]
   e. Matrix ---> [[],[],...]
   f. Tensor
   g. Dictionary ---> { key:value, key:value, ...]
   h. DataFrame
   i. GeoDataFrame
   j. NetCDF
   k. Spatial data indexing, spatial scan order or spatial data encoding

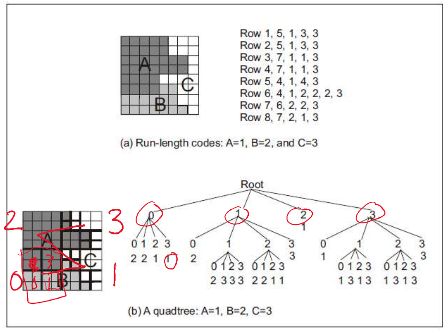


(a) row

(b) row-prime

(c) Morton

(D) Peano-Hilbret



Row 1, 5, 1, 3, 3
Row 2, 5, 1, 3, 3
Row 3, 7, 1, 1, 3
Row 4, 7, 1, 1, 3
Row 5, 4, 1, 4, 3
Row 6, 4, 1, 2, 2, 2, 3
Row 7, 6, 2, 2, 3
Row 8, 7, 2, 1, 3

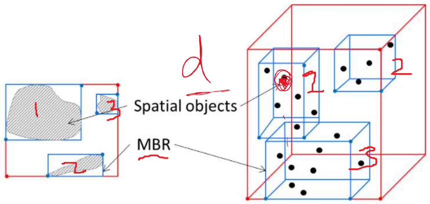
(a) Run-length codes: A=1, B=2, and C=3

(b) A quadtree: A=1, B=2, C=3

0	1
2	3

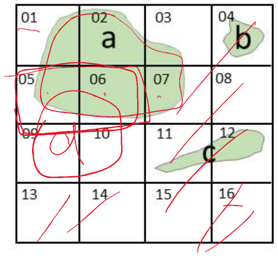
Yuan, M. (2010) Geographic Data Structures, Chapter 28 in Manual of Geospatial Science and Technology, Second Edition. Edited by John Bossler, James B. Campell, Robert McMaster, and Chris Rizos. P. 549-573. CRC Press. Boca Raton, Florida USA.

A more recent summary of spatial data indexing is available at https://gistbok.ucgis.org/bok-topics/spatial-indexing

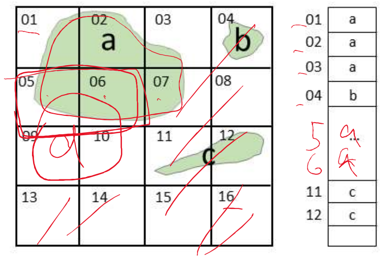


(a)

(b)



01	a
02	a
03	a
04	b
5	a
6	a
11	c
12	c



01	a
02	a
03	a
04	b
5	a
6	a
11	c
12	c

Figure 3. An example of a fixed grid structure.

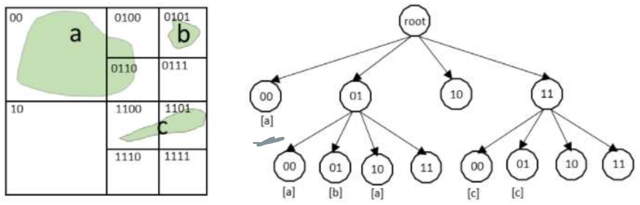


Figure 6. A representation of a quadtree structure.

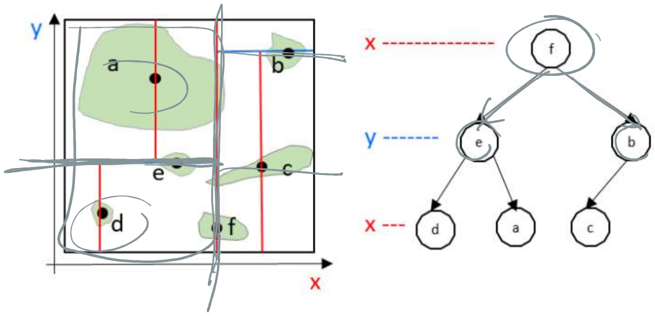


Figure 7. A representation of a KD-tree structure.

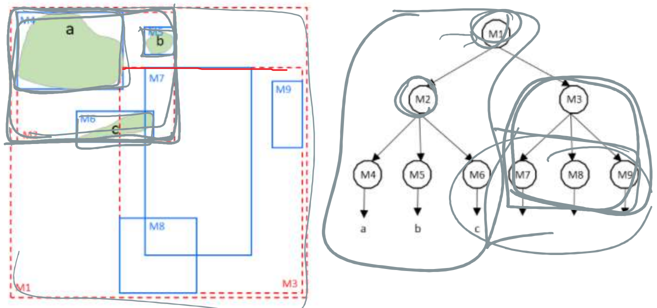
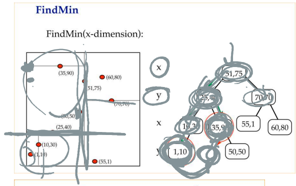
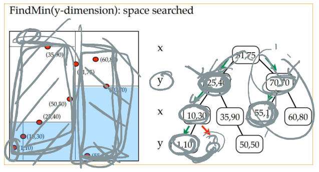


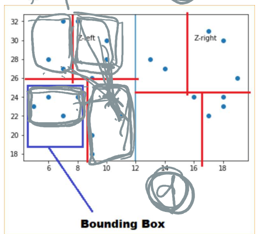
Figure 9. R-Tree Hierarchical Index in Minimum Bounding Rectangles (MBRs)

4. Location reference:
   a. EPSG code: spatial reference codes developed by European Petroleum Survey Group (EPSG): EPSG 4326 is WGS84. EPSG 3669 is Texas North Central in NAD83.
   b. Latitude and longitude
   c. Address
   d. Geohash code

FOSS4G



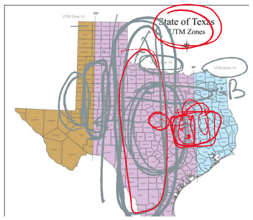




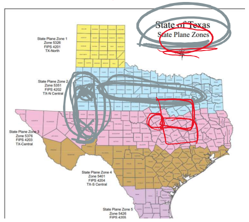
Spatial index creation in Quad-trees is faster as compared to R-trees.

R-trees are faster than Quad-trees for Nearest Neighbour queries while for window queries, Quad-trees are faster than R-trees.

m



ft £



geodetic

geodeeic

reference lines

4. Location reference.
    a. EPSG code: spatial reference codes developed by European Petroleum Survey Group (EPSG): EPSG 4326 is WGS84. EPSG 3669 is Texas North Central in NAD83.
    b. Latitude and longitude
    c. Address
    d. Geohash code

FOSS4G

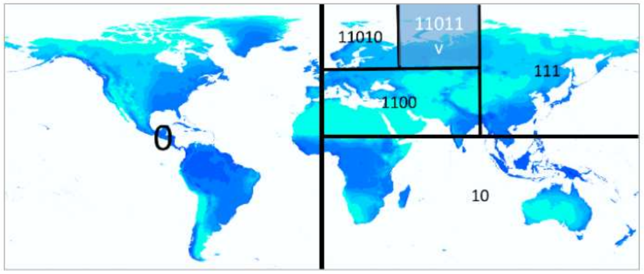
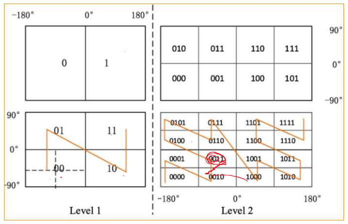


Figure 8. An example of a Geohash (with a hash code "v")



F What three words https://what3words.com/clip.apples.leap

5. Data representation
    a. Encoding: 1 - pass, 0 - fail
    b. Binary: 10 ->> 2, 100 ->> 4
    c. UTF-8: unicode transformation format - 8 bit (the default python encoding)
    d. Vector
    e. Raster
    f. TIN

Surface

deep

UTF 16

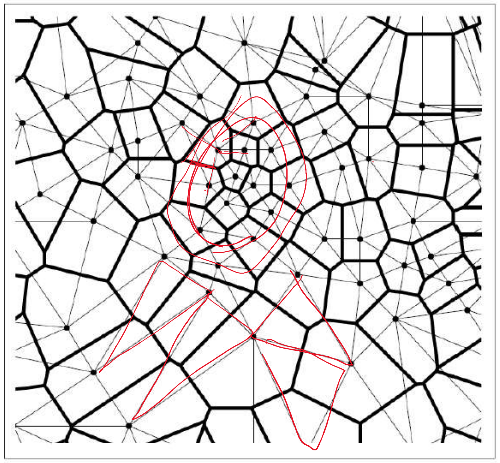
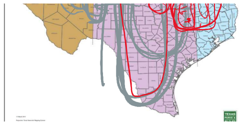
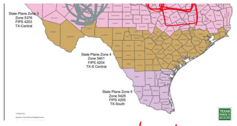
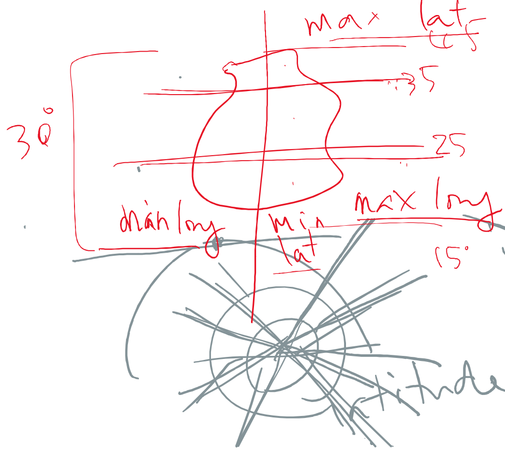
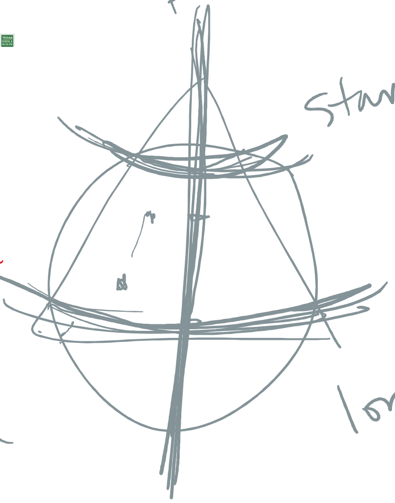


Figure 28.4: An example of Thiessen Polygons (thick lines) and Delauney Triangles (thin lines).









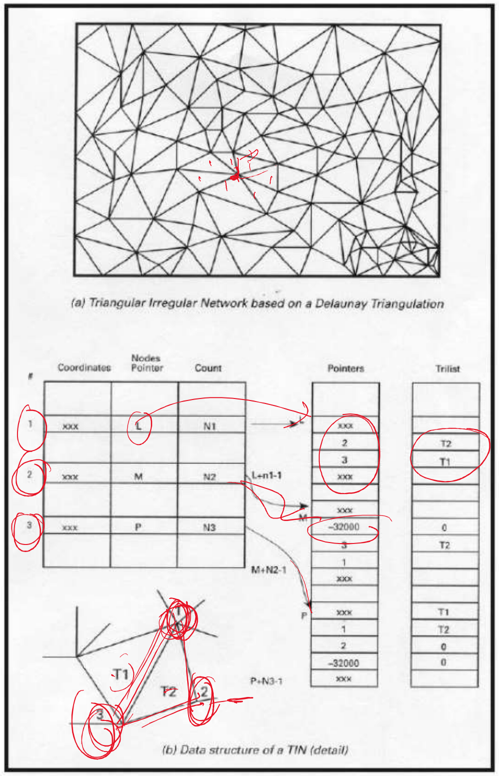
Delauney triangles

Thissen polygons

pit

peak

Grid



(a) Triangular Irregular Network based on a Delaunay Triangulation

(b) Data structure of a TIN (detail)

G. Network

6. Data model
   a. Relational: PostgreSQL (postGIS)
   b. Object-oriented
   c. Object-Relational: Geodatabase
   d. File: Shapefile, MongoDB (BSON files – similar to JSON files)

flat

coordinates

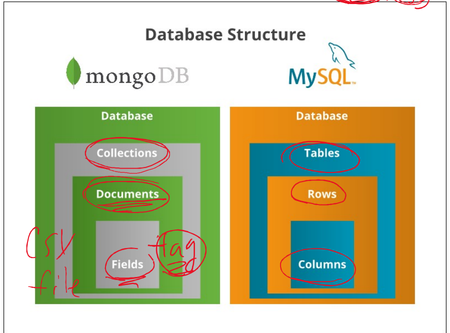
shp

dbf

shx

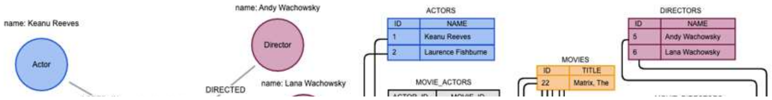
shx

prj

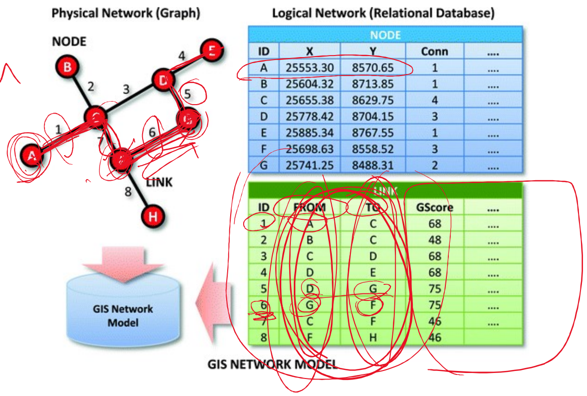


https://phoenixnap.com/kb/mongodb-vs-mysql

e. Graph: Neo4J-Spatial



stream



Logical Network (Relational Database)

NODE

ID	X	Y	Conn	....
A	25553.30	8570.65	1	....
B	25604.32	8713.85	1	....
C	25655.38	8629.75	4	....
D	25778.42	8704.15	3	....
E	25885.34	8767.55	1	....
F	25698.63	8558.52	3	....
G	25741.25	8488.31	2	....

LINK

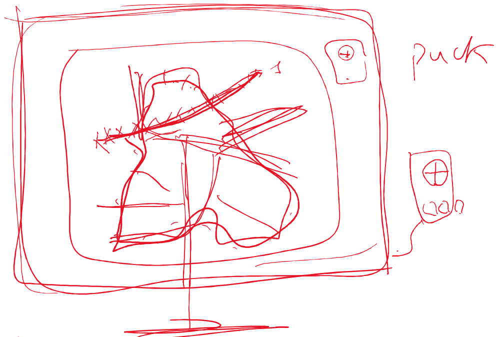
ID	FROM	TO	GScore	....
1	A	C	68	....
2	B	C	48	....
3	C	D	68	....
4	D	E	68	....
5	D	G	75	....
6	G	F	75	....
7	C	F	46	....
8	F	H	46	

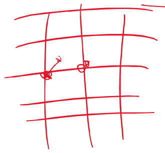
operator

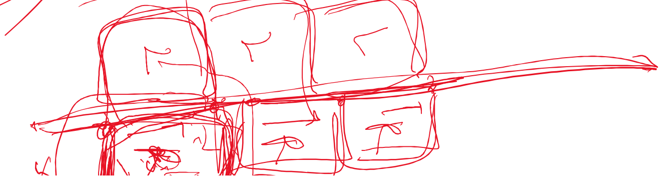
FOSS4G

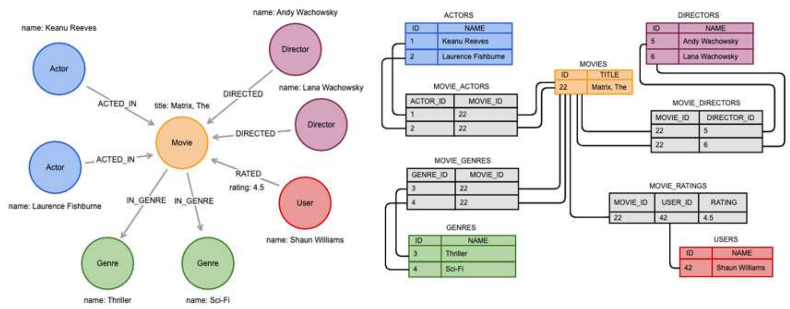
Left poly right poly

topological









ACTORS

ID	NAME
1	Keanu Reeves
2	Laurence Fishburne

DIRECTORS

ID	NAME
5	Andy Wachowsky
6	Lana Wachowsky

MOVIES

ID	TITLE
22	Matrix, The

MOVIE_ACTORS

ACTOR_ID	MOVIE_ID
1	22
2	22

MOVIE_DIRECTORS

MOVIE_ID	DIRECTOR_ID
22	5
22	6

MOVIE_GENRES

GENRE_ID	MOVIE_ID
3	22
4	22

MOVIE_RATINGS

MOVIE_ID	USER_ID	RATING
22	42	4.5

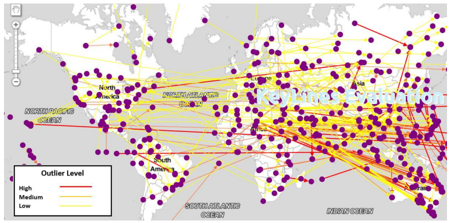
GENRES

ID	NAME
3	Thriller
4	Sci-Fi

USERS

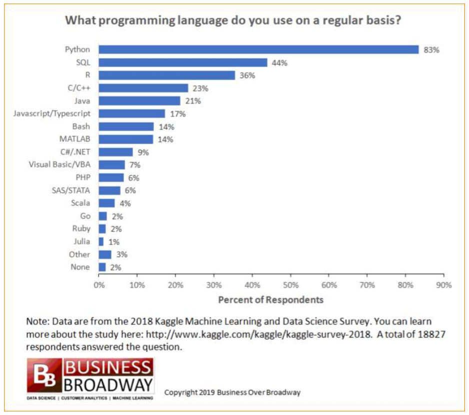
ID	NAME
42	Shaun Williams

https://www.profit4cloud.nl/blog/introduction-to-graph-databases-using-neo4j-part-1/



https://www.linkedin.com/pulse/finding-valuable-outliers-opportunities-using-graph-spatial-urich/

There are many programming languages for data science, but python is the most popular and versatile language for data science and data engineering. Hence, we will use python as the primary programming language in this class.



Note: Data are from the 2018 Kaggle Machine Learning and Data Science Survey. You can learn more about the study here: http://www.kaggle.com/kaggle/kaggle-survey-2018. A total of 18827 respondents answered the question.

Copyright 2019 Business Over Broadway