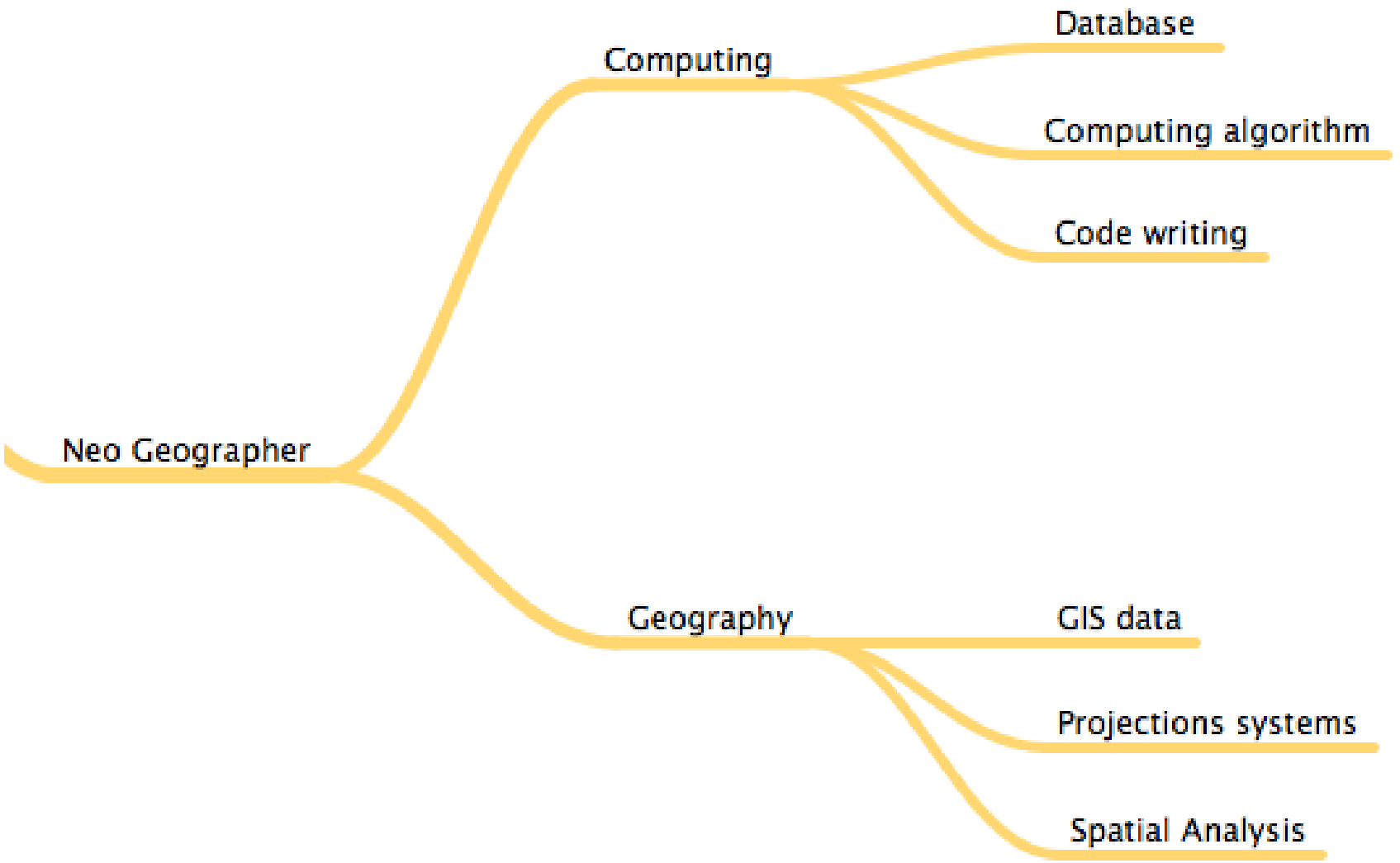
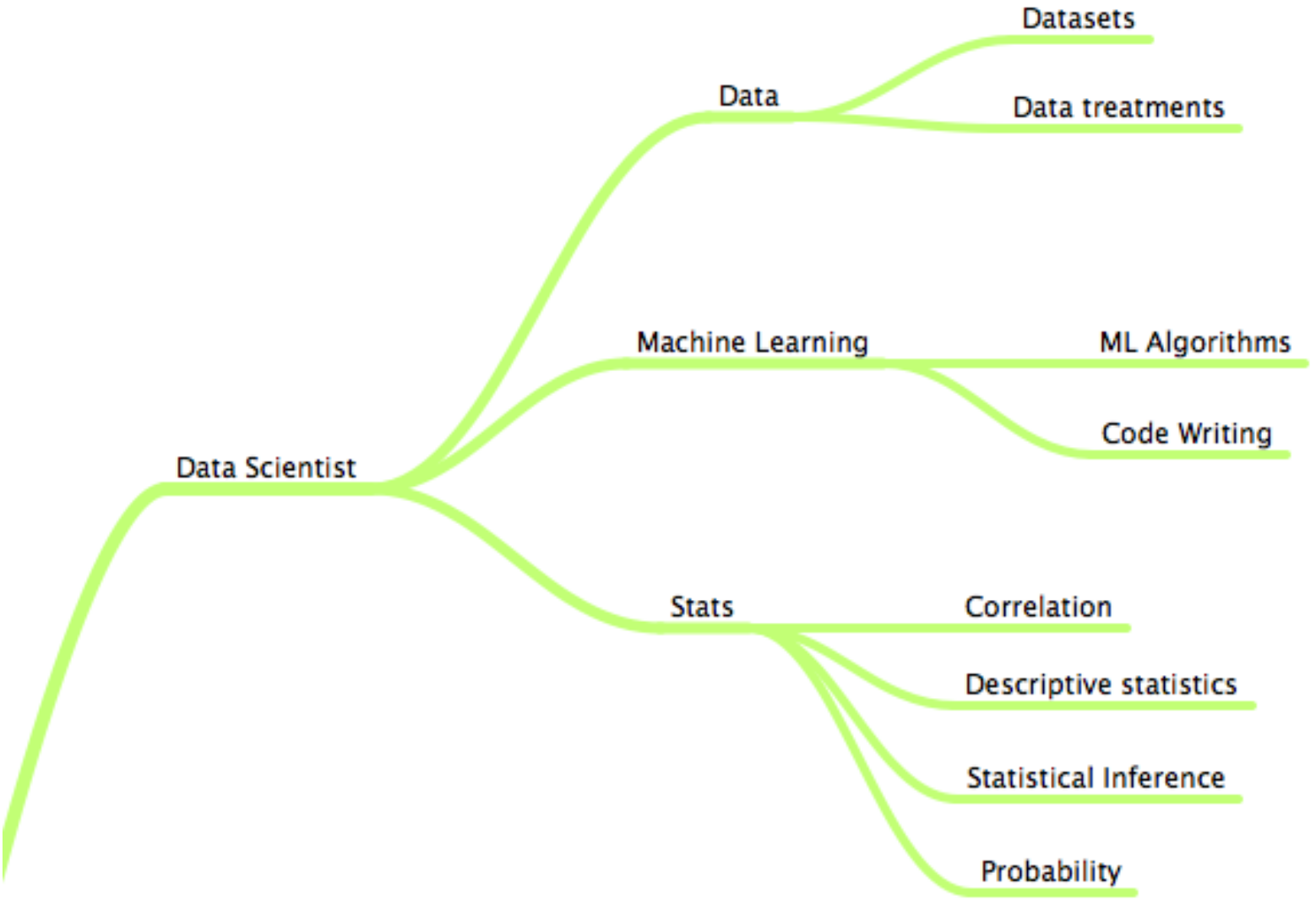
Les compétences du GéoDataScientist

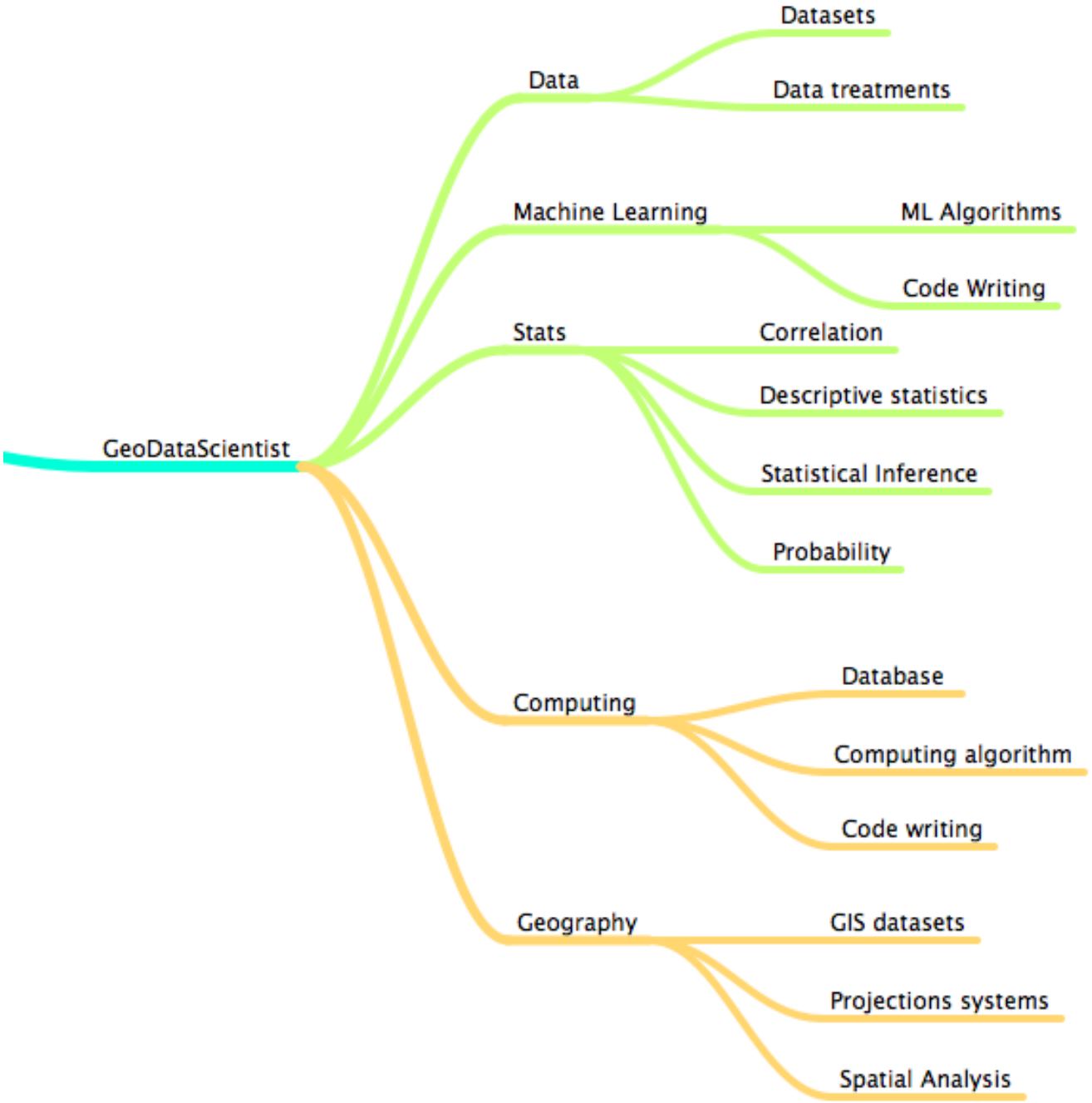
Decryptageo 2017

Neo Geographer

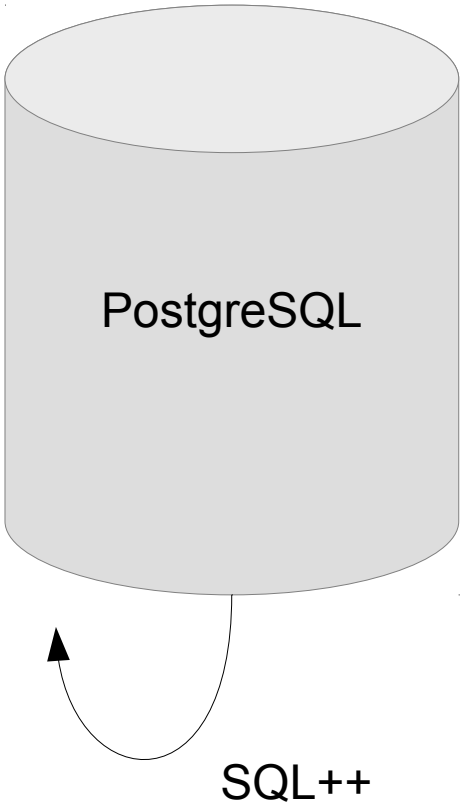

Neo Geographer
Computing
Database
Computing algorithm
Code writing
Geography
GIS data
Projections systems
Spatial Analysis

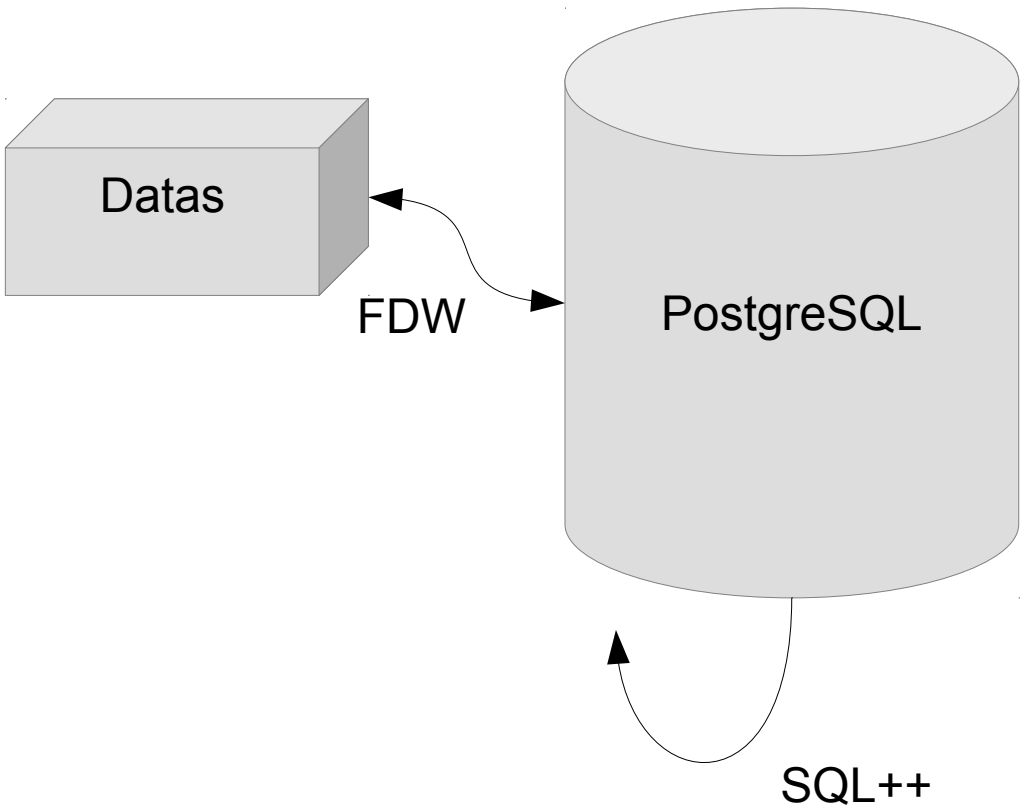
Data Scientist

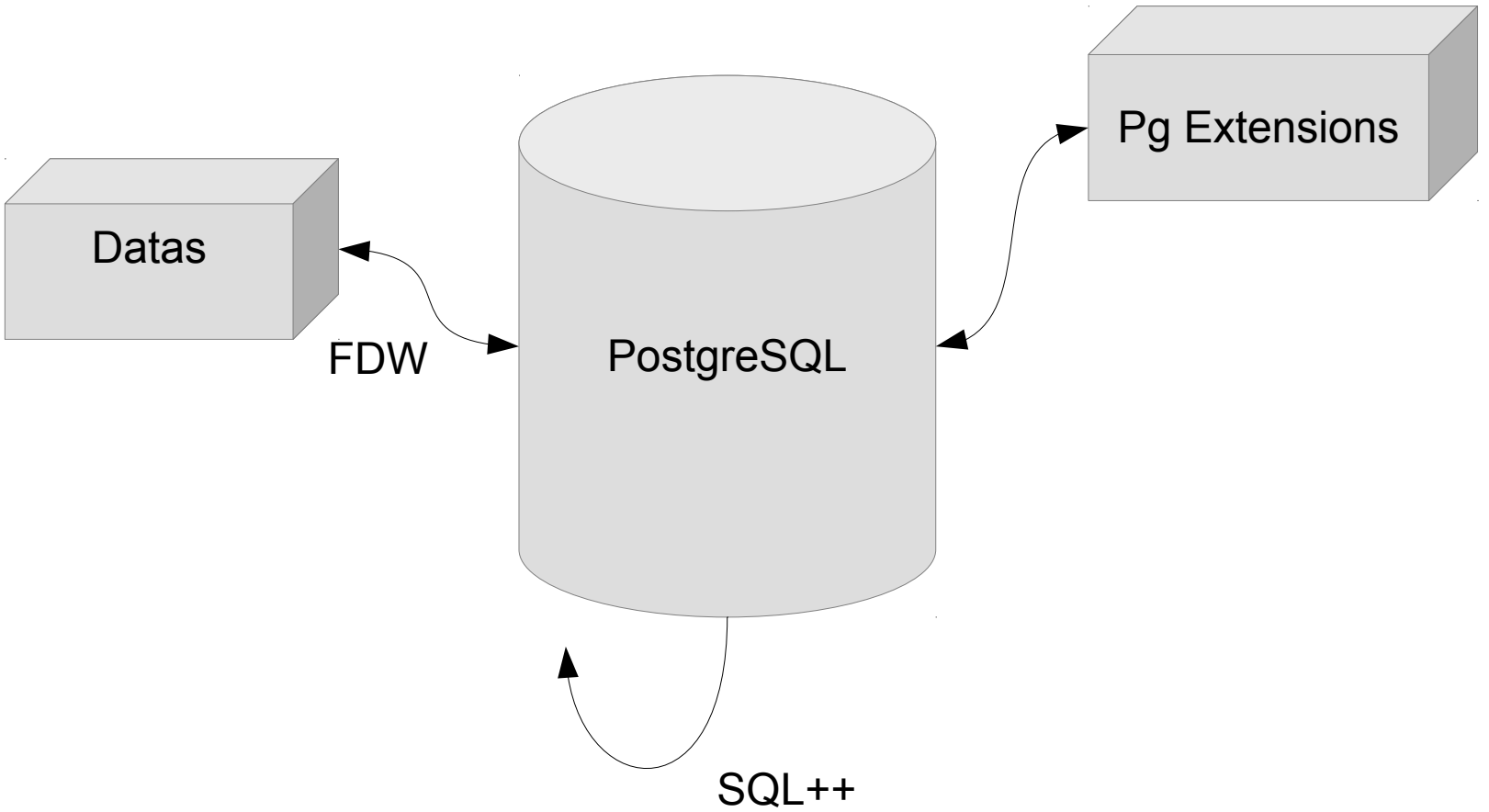

Data Scientist
Data
Datasets
Data treatments
Machine Learning
ML Algorithms
Code Writing
Stats
Correlation
Descriptive statistics
Statistical Inference
Probability

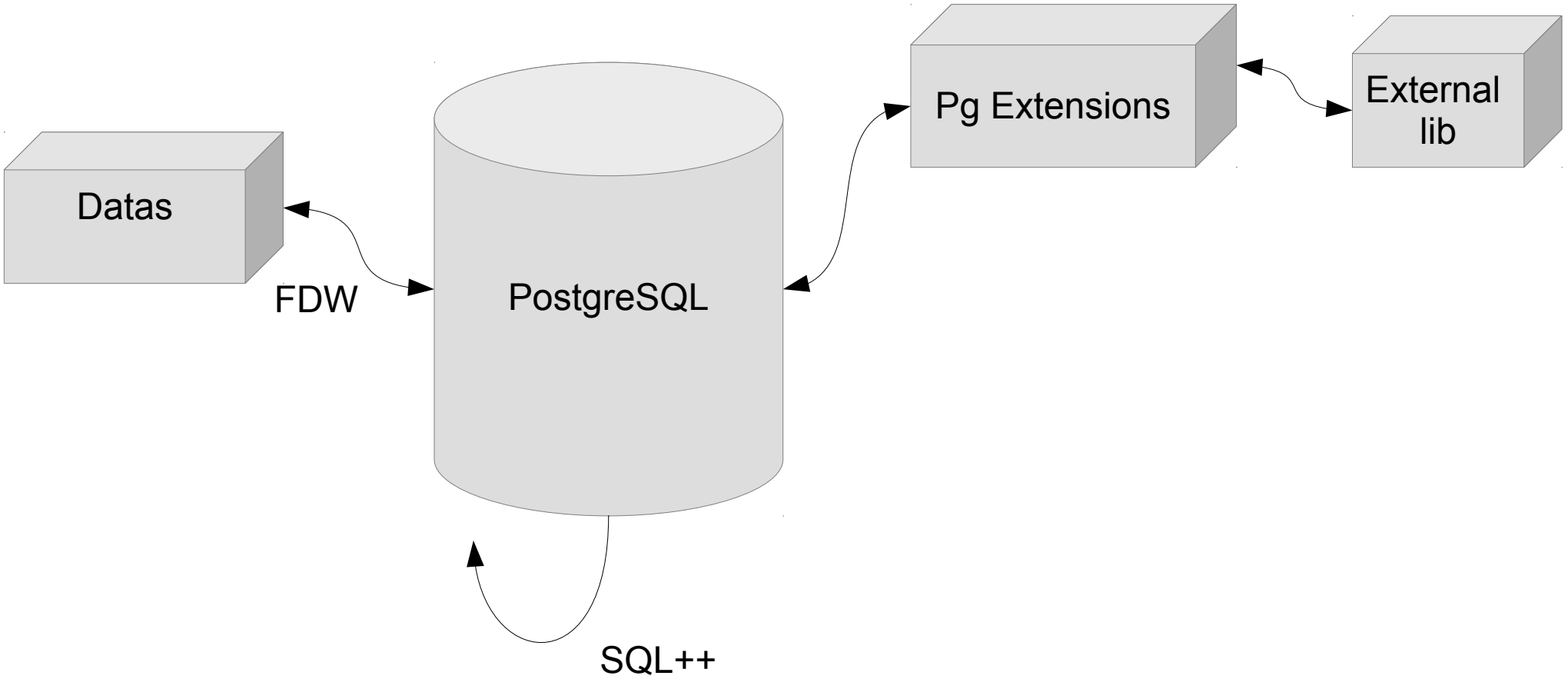

GeoDataScientist
Data
Datasets
Data treatments
Machine Learning
ML Algorithms
Code Writing
Stats
Correlation
Descriptive statistics
Statistical Inference
Probability
Computing
Database
Computing algorithm
Code writing
Geography
GIS datasets
Projections systems
Spatial Analysis

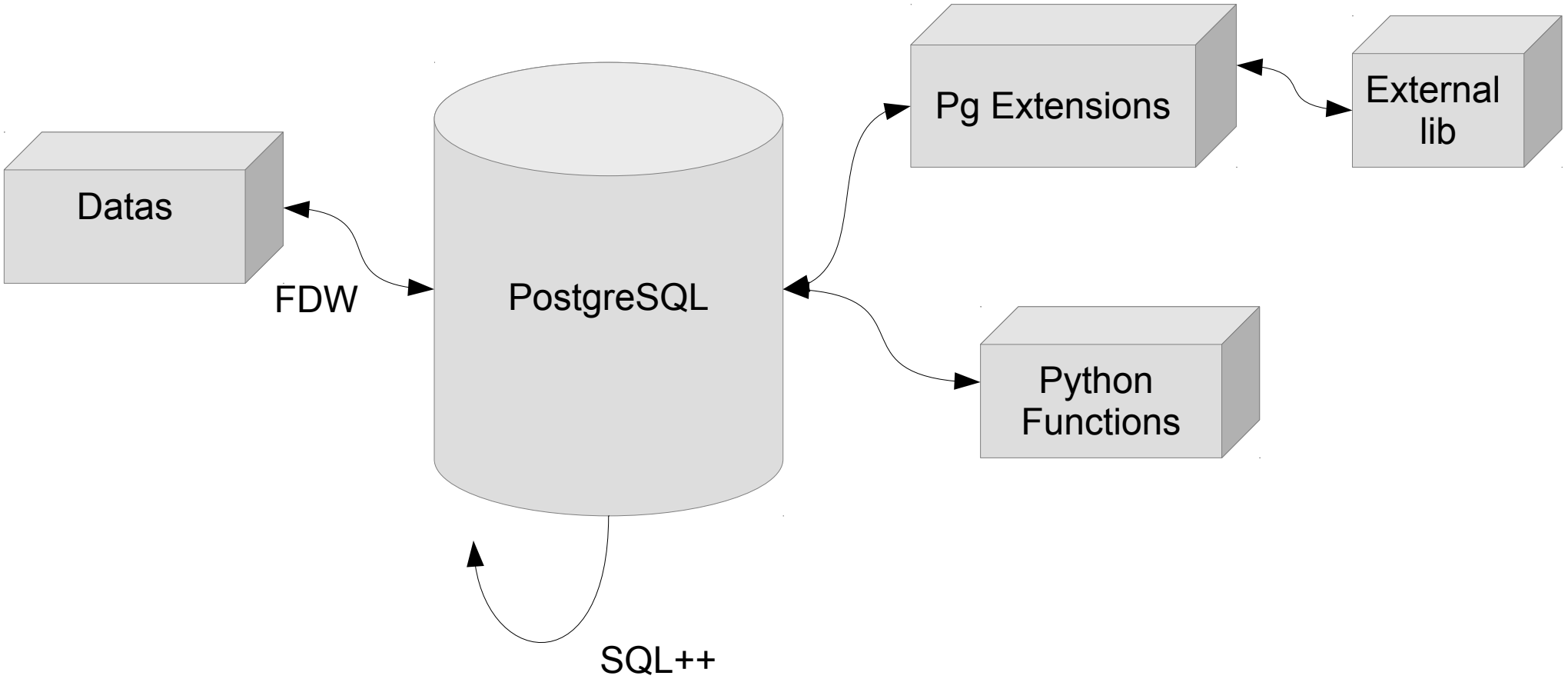
Thick Database


PostgreSQL
SQL++


Datas
FDW
PostgreSQL
SQL++


Datas
FDW
PostgreSQL
Pg Extensions
SQL++


Datas
FDW
PostgreSQL
Pg Extensions
External
lib
SQL++

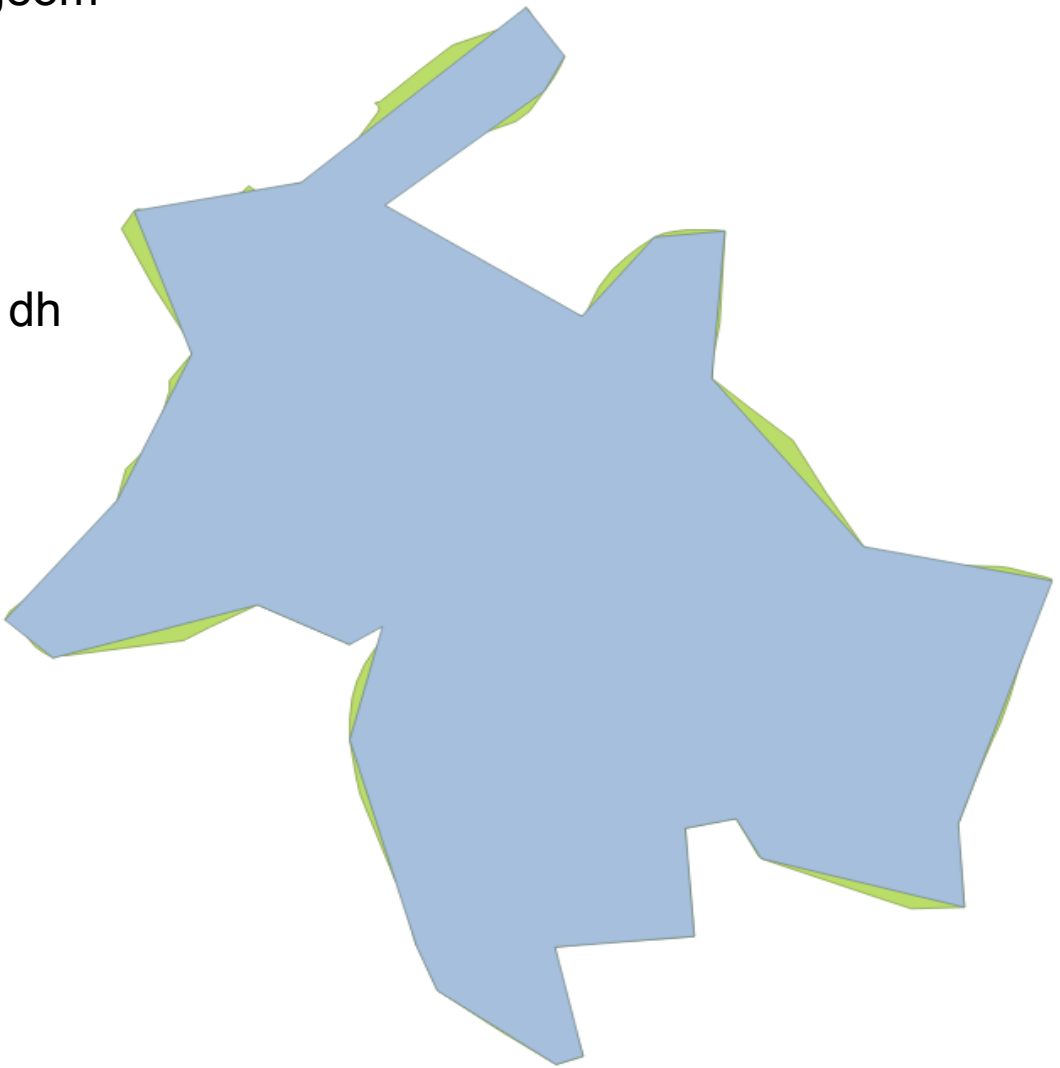

Datas
FDW
PostgreSQL
Pg Extensions
External lib
Python Functions
SQL++

" Everything is related to everything else,
but near things are more related than distant things. "

W. Tobler

ST_HausdorffDistance

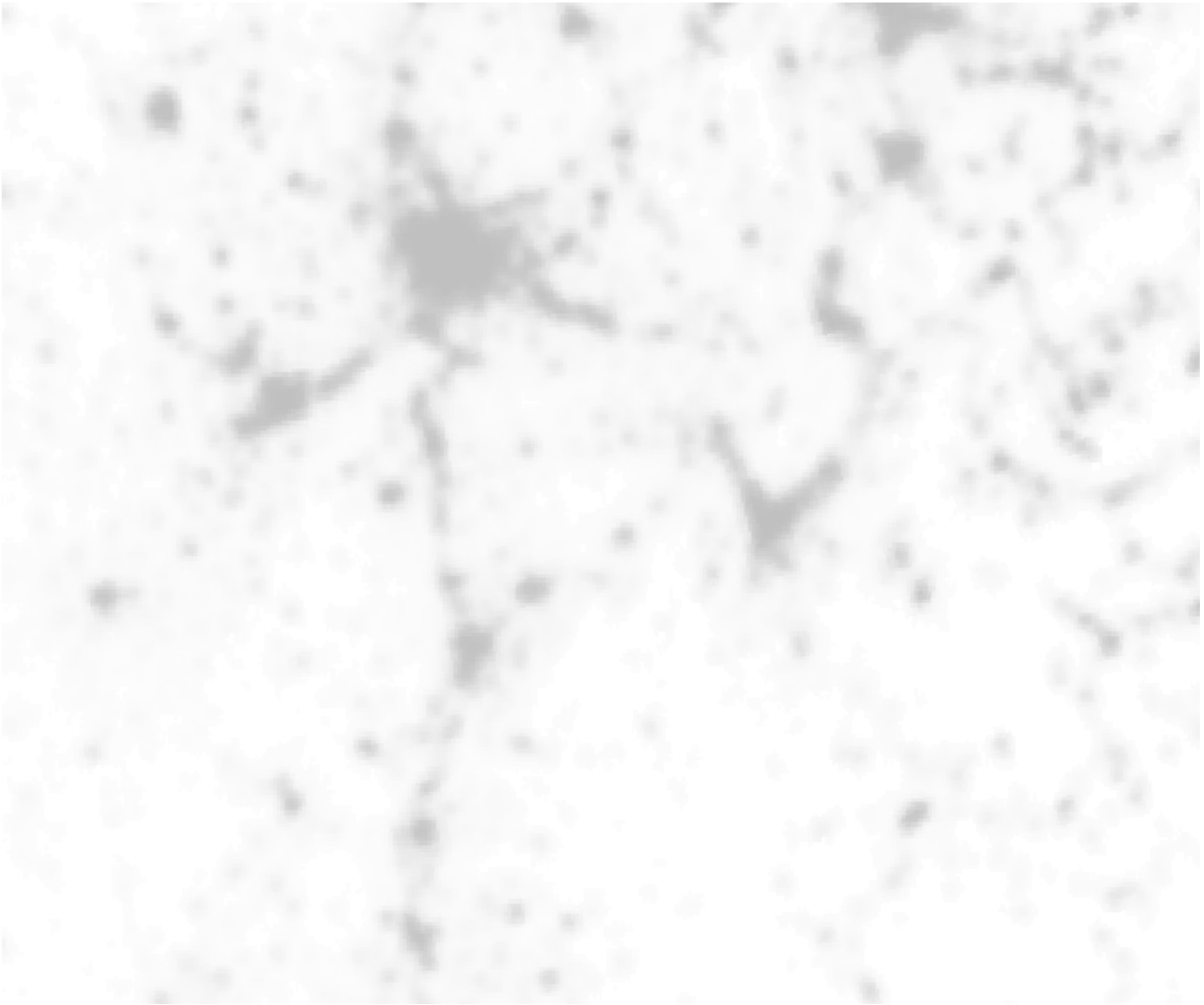
```
WITH a AS (
  SELECT id, ST_Simplify(geom, 5000) AS geom
  FROM own.commune
)

SELECT a.id, b.id,
ST_HausdorffDistance(a.geom, b.geom) AS dh
FROM a, own.commune b
WHERE nom_com = 'Lyon'
ORDER BY dh ASC
LIMIT 5;
```



```
 id  |  id  |        dh
------+------+------------------
 1347 | 1347 | 185.139093997864
 1072 | 1347 | 6681.60493070321
 2461 | 1347 | 6817.89817025694
 2824 | 1347 | 7149.21791806655
  344 | 1347 | 7929.70883765602
```

But, could we get a bit deeper in our (spatial) analysis ?

Light Pollution @Night

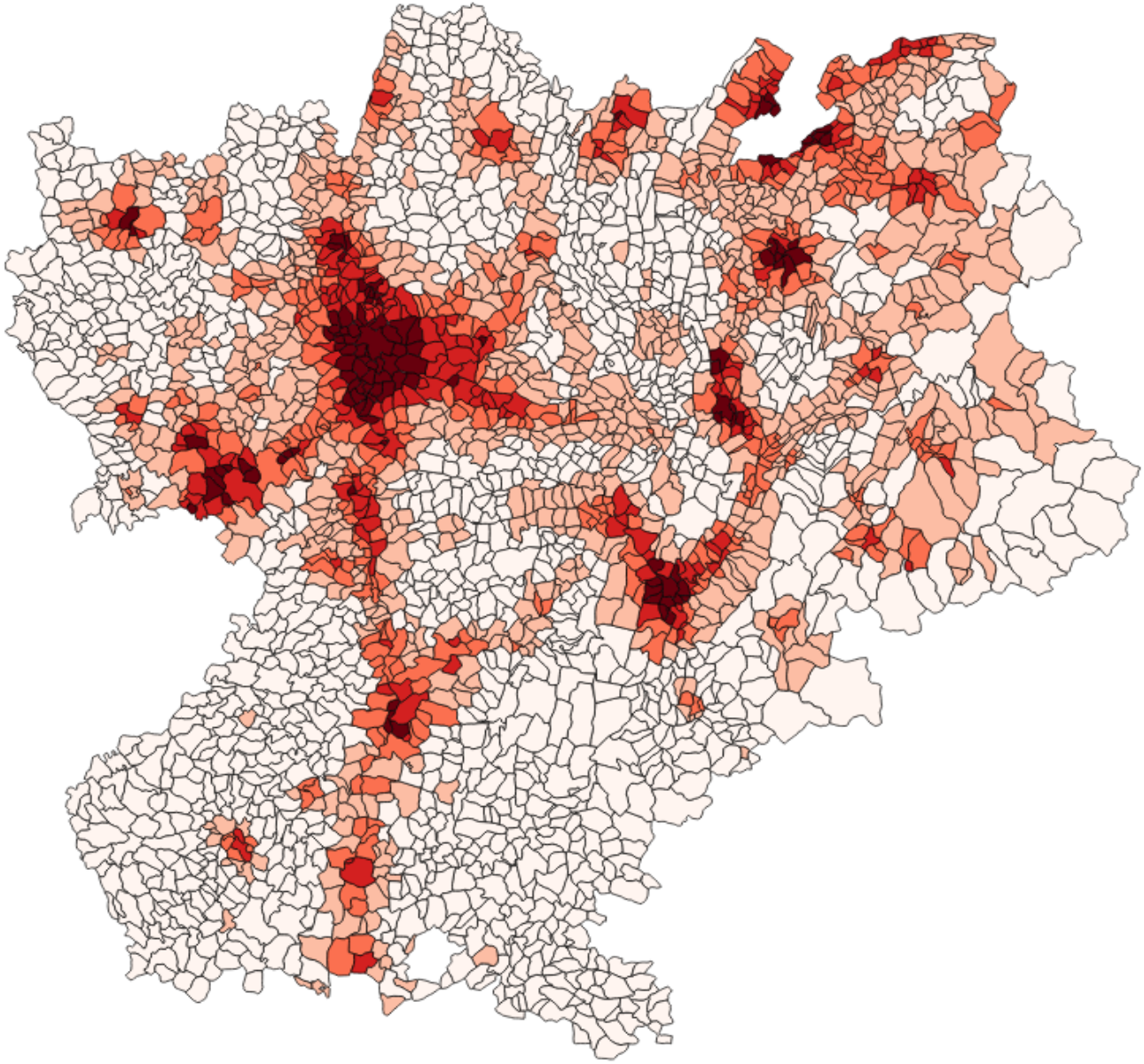


Open Data from : http://geodata.grid.unep.ch - 2003 Raster

Raster (light pollution) / Vector (area) Intersection

```
WITH ln AS
(
   SELECT id, avg(px) AS light

   FROM
   (
      SELECT  id,  ST_Value(rast, ST_SetSrid((ST_Dumppoints(pts)).geom, 2154)) AS px
      FROM (
                  SELECT id, geom AS pts  FROM own.commune
               ) AS t , r
      WHERE ST_Intersects(rast, pts)
    ) AS tt

   GROUP BY id
)

UPDATE own.commune c SET light = ln.light_pollution FROM ln WHERE c.id = ln.id
```

Light pollution by area



Road density by area

```
ALTER TABLE own.commune ADD COLUMN road_density_2016 numeric;


WITH rd AS (

	SELECT c.id,
	   (SUM(ST_Length(  ST_Intersection(c.geom, r.geom)))   /    ST_Area(c.geom)) AS road_density
	FROM own.commune c, osm.roads_2016 r
	WHERE ST_Intersects(c.geom, r.geom)
	GROUP BY c.id

)

UPDATE own.commune c SET road_density_2016 = rd.road_density FROM rd WHERE c.id = rd.id
```

Table 9-50. Aggregate Functions for Statistics

Function	Argument Type	Return Type	Description
corr(*Y*, *X*)	double precision	double precision	correlation coefficient
covar_pop(*Y*, *X*)	double precision	double precision	population covariance
covar_samp(*Y*, *X*)	double precision	double precision	sample covariance
regr_avgx(*Y*, *X*)	double precision	double precision	average of the independent variable (sum(*X*)/*N*)
regr_avgy(*Y*, *X*)	double precision	double precision	average of the dependent variable (sum(*Y*)/*N*)
regr_count(*Y*, *X*)	double precision	bigint	number of input rows in which both expressions are nonnull
regr_intercept(*Y*, *X*)	double precision	double precision	y-intercept of the least-squares-fit linear equation determined by the (*X*, *Y*) pairs
regr_r2(*Y*, *X*)	double precision	double precision	square of the correlation coefficient
regr_slope(*Y*, *X*)	double precision	double precision	slope of the least-squares-fit linear equation determined by the (*X*, *Y*) pairs
regr_sxx(*Y*, *X*)	double precision	double precision	sum(*X*^2) - sum(*X*)^2/*N* ("sum of squares" of the independent variable)
regr_sxy(*Y*, *X*)	double precision	double precision	sum(*X***Y*) - sum(*X*) * sum(*Y*)/*N* ("sum of products" of independent times dependent variable)
regr_syy(*Y*, *X*)	double precision	double precision	sum(*Y*^2) - sum(*Y*)^2/*N* ("sum of squares" of the dependent variable)
stddev(*expression*)	smallint, int, bigint, real, double precision, or numeric	double precision for floating-point arguments, otherwise numeric	historical alias for stddev_samp
stddev_pop(*expression*)	smallint, int, bigint, real, double precision, or numeric	double precision for floating-point arguments, otherwise numeric	population standard deviation of the input values
stddev_samp(*expression*)	smallint, int, bigint, real, double precision, or numeric	double precision for floating-point arguments, otherwise numeric	sample standard deviation of the input values
variance(*expression*)	smallint, int, bigint, real, double precision, or numeric	double precision for floating-point arguments, otherwise numeric	historical alias for var_samp
var_pop(*expression*)	smallint, int, bigint, real, double precision, or numeric	double precision for floating-point arguments, otherwise numeric	population variance of the input values (square of the population standard deviation)
var_samp(*expression*)	smallint, int, bigint, real, double precision, or numeric	double precision for floating-point arguments, otherwise numeric	sample variance of the input values (square of the sample standard deviation)

SELECT corr (pop_density, light)::numeric(4,4) FROM own.commune;

0.6533

- - OSM 08/2014
SELECT corr (road_density, light)::numeric(4,4) FROM own.commune;

0.7573

- - OSM 08/2016
SELECT corr (road_density, light)::numeric(4,4) FROM own.commune;

0.7782

" Everything is related to everything else,
but near things are more related than distant things. "

W. Tobler

Moran I - Spatial Autocorrelation Coefficient

1 → Strong Spatial Correlation

0 → Random

-1 → Perfectly dispersed

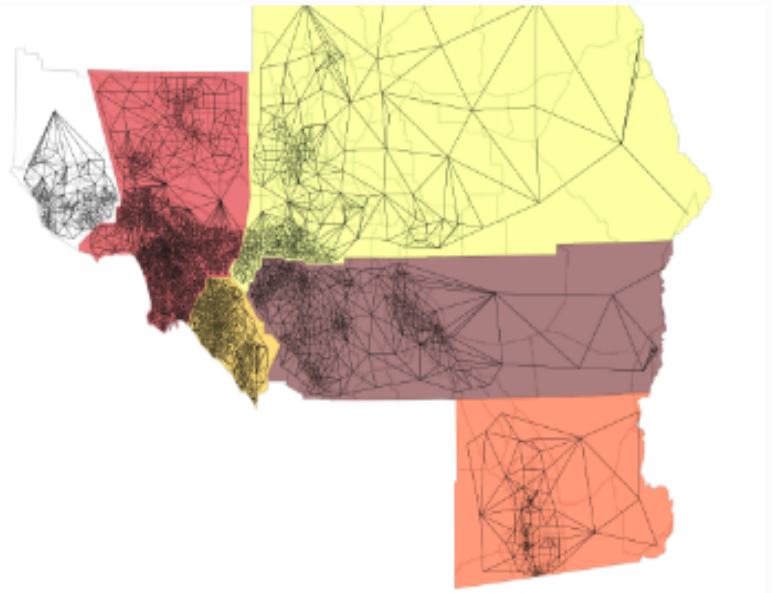
$$I = \frac{N}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(X_i - \bar{X})(X_j - \bar{X})}{\sum_i (X_i - \bar{X})^2}$$

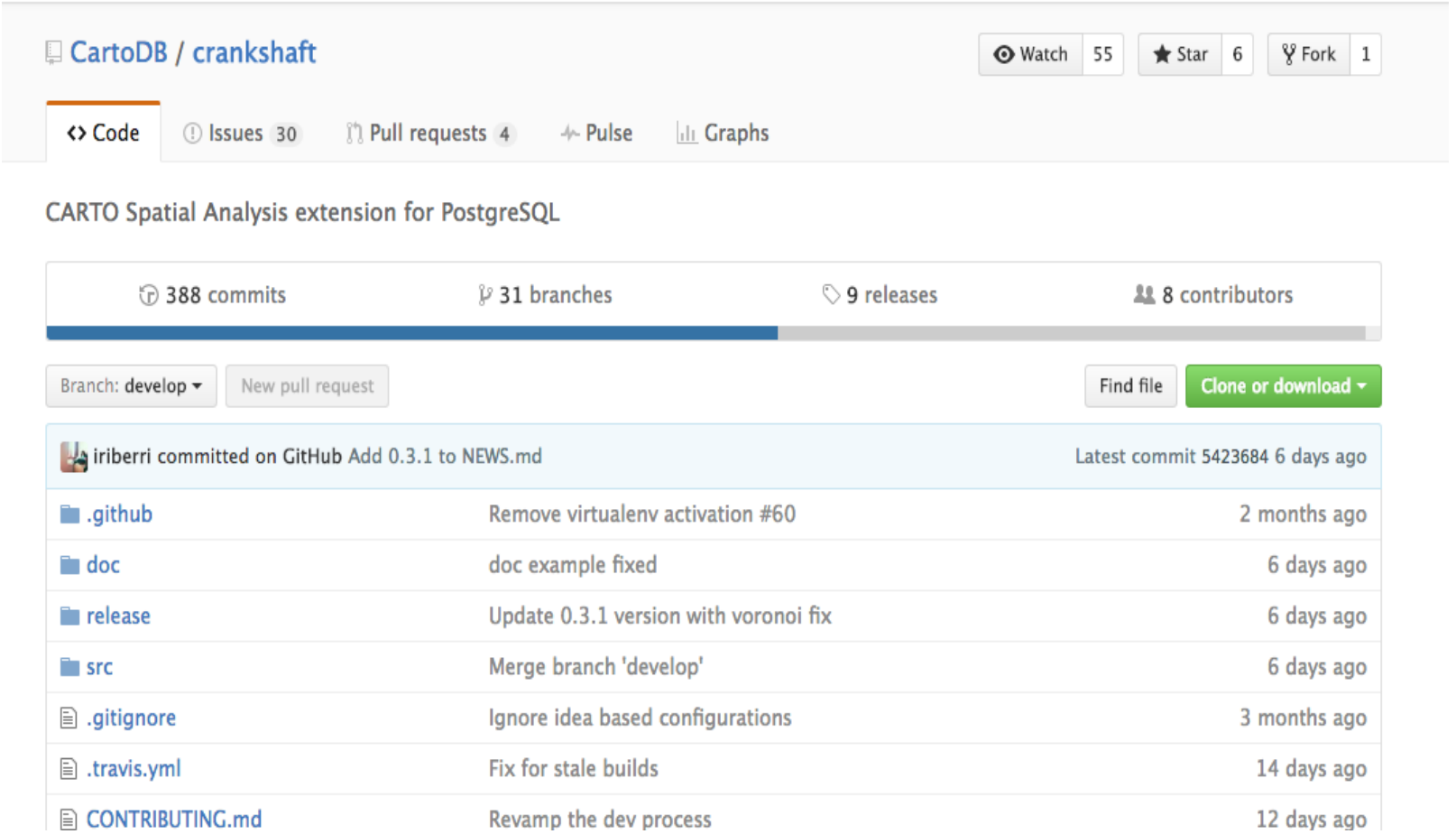
Humm, do we really need R ?

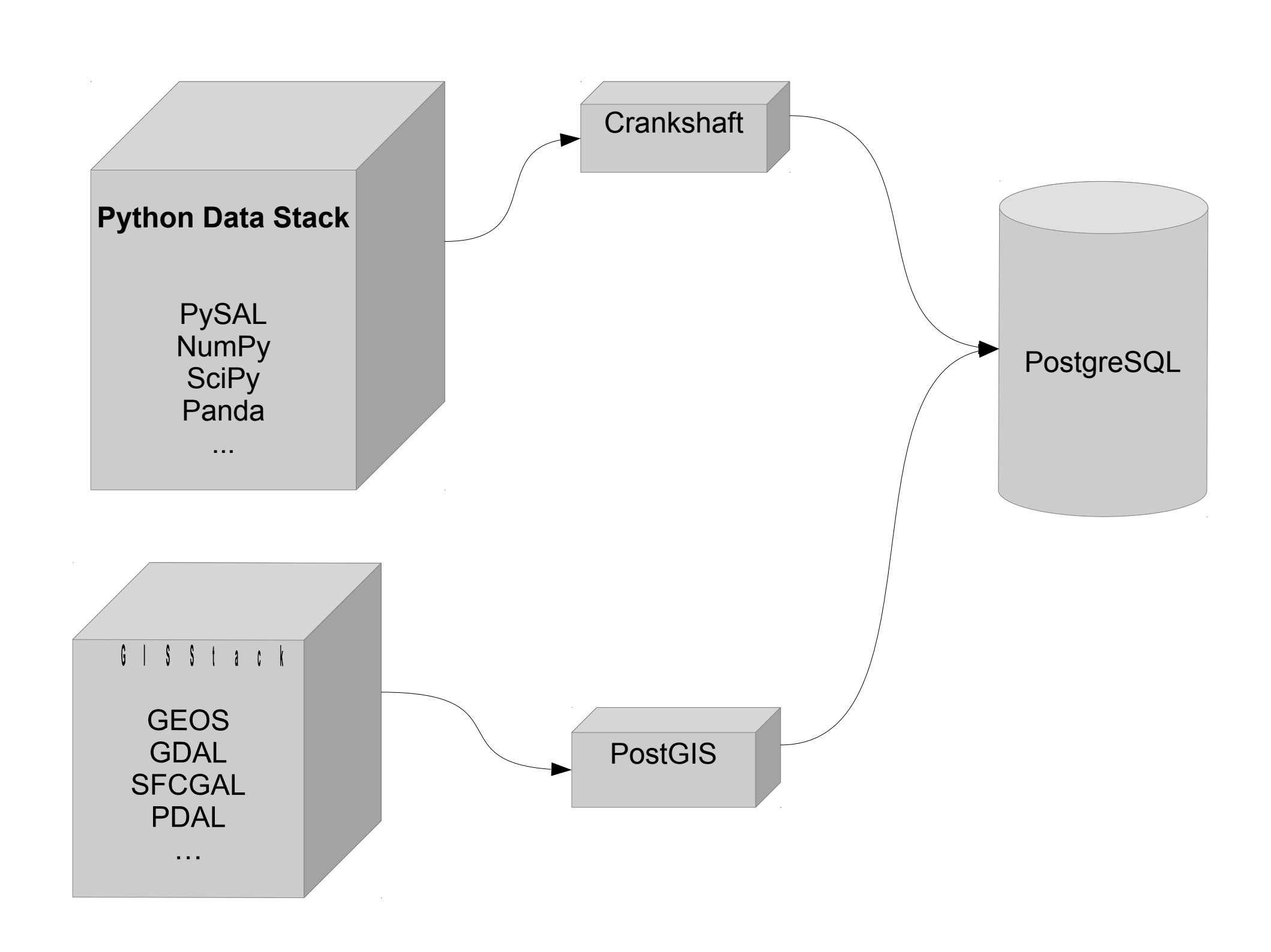
http://pysal.github.io/grid.html

PySAL: Python Spatial Analysis Library

This page collects links to examples using pysal. Click on each figure to see access the full example with code included.




CartoDB / crankshaft
Watch 55
Star 6
Fork 1
Code
Issues 30
Pull requests 4
Pulse
Graphs
CARTO Spatial Analysis extension for PostgreSQL
388 commits
31 branches
9 releases
8 contributors
Branch: develop
New pull request
Find file
Clone or download
iriberri committed on GitHub Add 0.3.1 to NEWS.md
Latest commit 5423684 6 days ago
.github Remove virtualenv activation #60 2 months ago
doc doc example fixed 6 days ago
release Update 0.3.1 version with voronoi fix 6 days ago
src Merge branch 'develop' 6 days ago
.gitignore Ignore idea based configurations 3 months ago
.travis.yml Fix for stale builds 14 days ago
CONTRIBUTING.md Revamp the dev process 12 days ago


Python Data Stack
PySAL
NumPy
SciPy
Panda
...
Crankshaft
PostgreSQL
GIS Stack
GEOS
GDAL
SFCGAL
PDAL
...
PostGIS

```
SELECT moran::numeric(10, 4)

FROM  cdb_crankshaft.cdb_areasofinterestGlobal(

        'SELECT * FROM own.commune',  - - data table
        'light',    - - column name to check
        'knn',      - - weight : queen or knn
        5,          - - k value (for knn)

        99, 'geom', 'id'
    )
```

queen	0.8235
knn5	0.8201
knn20	0.6687
knn50	0.5220

```
WITH m AS (

  SELECT aoi.*, c.id, c.nom_com , c.geom

  FROM cdb_crankshaft.cdb_areasofinterestlocal(
                    'SELECT * FROM own.commune',
                    'light',
                    'knn',
                    5,
                    99,
                    'geom',
                    'id') As aoi
  JOIN own.commune As c
  ON c.id = aoi.rowid

)

SELECT quads, geom, ow_number() OVER() AS id
FROM u
WHERE quads = 'HH' OR quads = 'LL'
```

Others OSM data mining initiatives

- Contributors Metadata :

  https://agile-online.org/conference_paper/cds/agile_2013/short_papers/sp_s4.2_arsanjani.pdf

- Socio-Economical OpenData Intersection :

  http://www.ciesin.org/binaries/web/global/news/2016/bogdan_cirlugea_osm_validation_v2.pdf

http://data.grandlyon.com/

HISTORIQUE DES DISPONIBILITÉS DES STATIONS VÉLO'V

La donnée historique des disponibilités des stations Vélo'V propose un historique sur les 7 derniers jours au pas de 5 minutes. Cette donnée dispose :

- Des services WMS, WFS, KML, SHP standards (idem stations Vélov)
- Du service SOS pour les requêtes temporelles sur la disponibilité des vélos et des bornettes
- Du graphe SOS sur la disponibilité des vélos et des bornettes sur les 7 derniers jours
- De la visionneuse temporelle WMS-T au pas de 5 minutes.

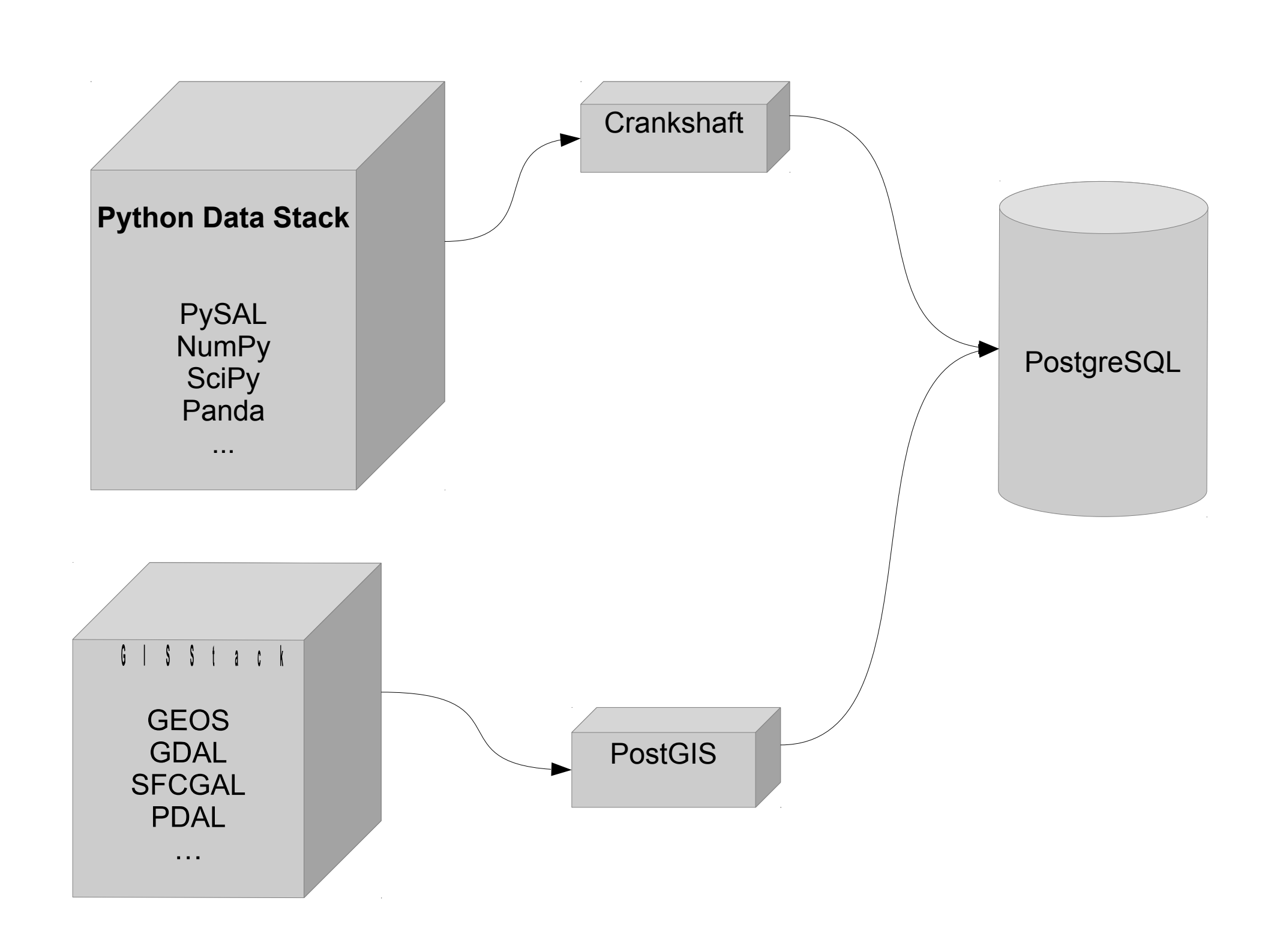
+ En savoir plus

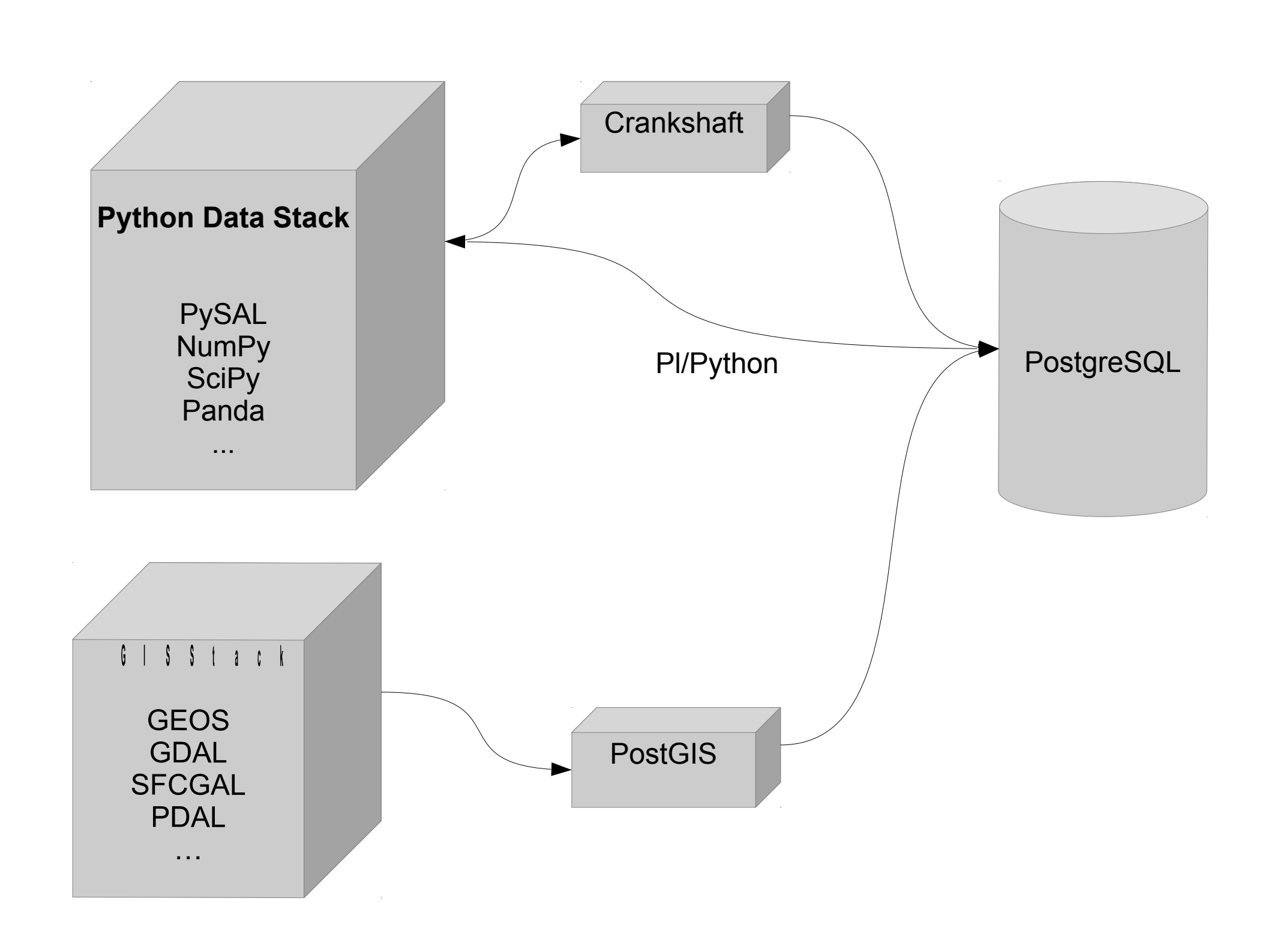
```
WITH a AS (
    SELECT *,
           bikes::numeric / (bikes + stands)::numeric AS avl,
           extract(hour FROM timestamp) h
    FROM gl.velov
    WHERE extract(isodow FROM timestamp) BETWEEN 1 AND 5
      AND (bikes + stands) != 0
  )

SELECT 'Monday to Friday' AS days, count(*) FROM a
UNION
SELECT 'Monday to Sunday' AS days, count(*)  FROM gl.velov
```

```
       days              |  count
-------------------------+---------
 Monday to Friday   | 2118859
 Monday to Sunday | 2830145
```

```
WITH a AS (
   SELECT *, bikes::numeric / (bikes + stands)::numeric AS avl, extract(hour FROM timestamp) h
   FROM gl.velov
   WHERE extract(isodow FROM timestamp) BETWEEN 1 AND 5 AND (bikes + stands) != 0
 ), b AS ( SELECT station id, h, avg(avl) AS avl FROM a GROUP BY id, h)
 , c1 AS ( SELECT id, h, avl FROM b WHERE id = 2012)
 , c2 AS ( SELECT id, h, avl FROM b WHERE id = 3001)
 , c3 AS ( SELECT id, h, avl FROM b WHERE id = 1002)
 , c4 AS ( SELECT id, h, avl FROM b WHERE id = 2035)

SELECT c1.h,
            c1.avl::numeric(10,3) AS Bellecour,
           c2.avl::numeric(10, 3) AS PartDieu,
           c3.avl::numeric(10, 3) AS Opera,
           c4.avl::numeric(10, 3) AS Republique
FROM c1, c2, c3, c4
WHERE c1.h = c2.h AND c2.h = c3.h AND c3.h = c4.h
ORDER BY h
```

h	bellecour	partdieu	opera	republique
0	0.775	0.042	0.349	0.361
1	0.744	0.065	0.285	0.307
2	0.731	0.062	0.158	0.223
3	0.725	0.068	0.136	0.235
4	0.732	0.107	0.123	0.201
5	0.747	0.238	0.100	0.177
6	0.749	0.651	0.158	0.157
7	0.545	0.847	0.143	0.176
8	0.256	0.786	0.126	0.315
9	0.306	0.815	0.286	0.692
10	0.424	0.876	0.356	0.843
11	0.510	0.839	0.322	0.835
12	0.635	0.801	0.371	0.840
13	0.737	0.793	0.398	0.902
14	0.598	0.858	0.470	0.886
15	0.611	0.861	0.531	0.816
16	0.712	0.882	0.555	0.777
17	0.586	0.854	0.692	0.790
18	0.671	0.519	0.768	0.627
19	0.744	0.161	0.829	0.418
20	0.790	0.042	0.856	0.516
21	0.840	0.042	0.913	0.690
22	0.796	0.049	0.792	0.656
23	0.794	0.066	0.568	0.551


Python Data Stack
PySAL
NumPy
SciPy
Panda
...
Crankshaft
PostgreSQL
GIS Stack
GEOS
GDAL
SFCGAL
PDAL
...
PostGIS


Python Data Stack
PySAL
NumPy
SciPy
Panda
...
Crankshaft
PostgreSQL
Pl/Python
GIS Stack
GEOS
GDAL
SFCGAL
PDAL
...
PostGIS

docs.scipy.org/doc/scipy/reference/signal.html

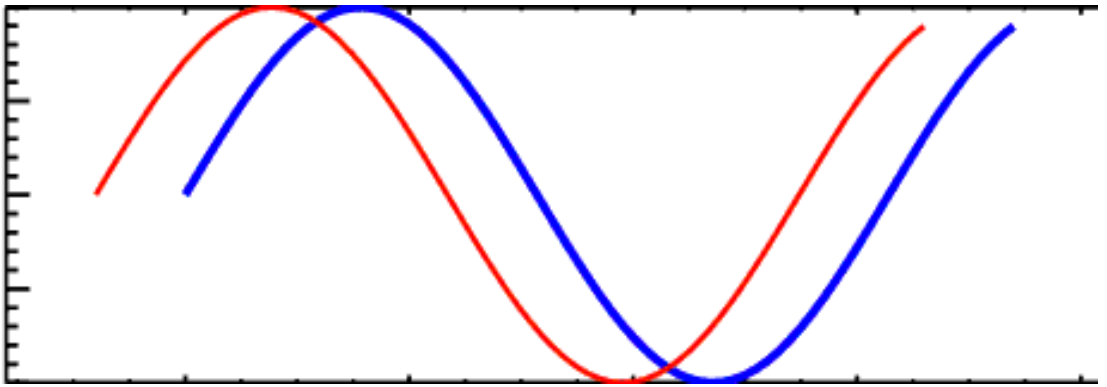
SciPy.org Sponsored By ENTHOUGHT

Scipy.org | Docs | SciPy v0.18.1 Reference Guide

Signal processing (scipy.signal)

Convolution

convolve(in1, in2[, mode])	Convolve two N-dimensional arrays.
correlate(in1, in2[, mode])	Cross-correlate two N-dimensional arrays.
fftconvolve(in1, in2[, mode])	Convolve two N-dimensional arrays using FFT.
convolve2d(in1, in2[, mode, boundary, fillvalue])	Convolve two 2-dimensional arrays.
correlate2d(in1, in2[, mode, boundary, ...])	Cross-correlate two 2-dimensional arrays.
sepfir2d((input, hrow, hcol) -> output)	Description:

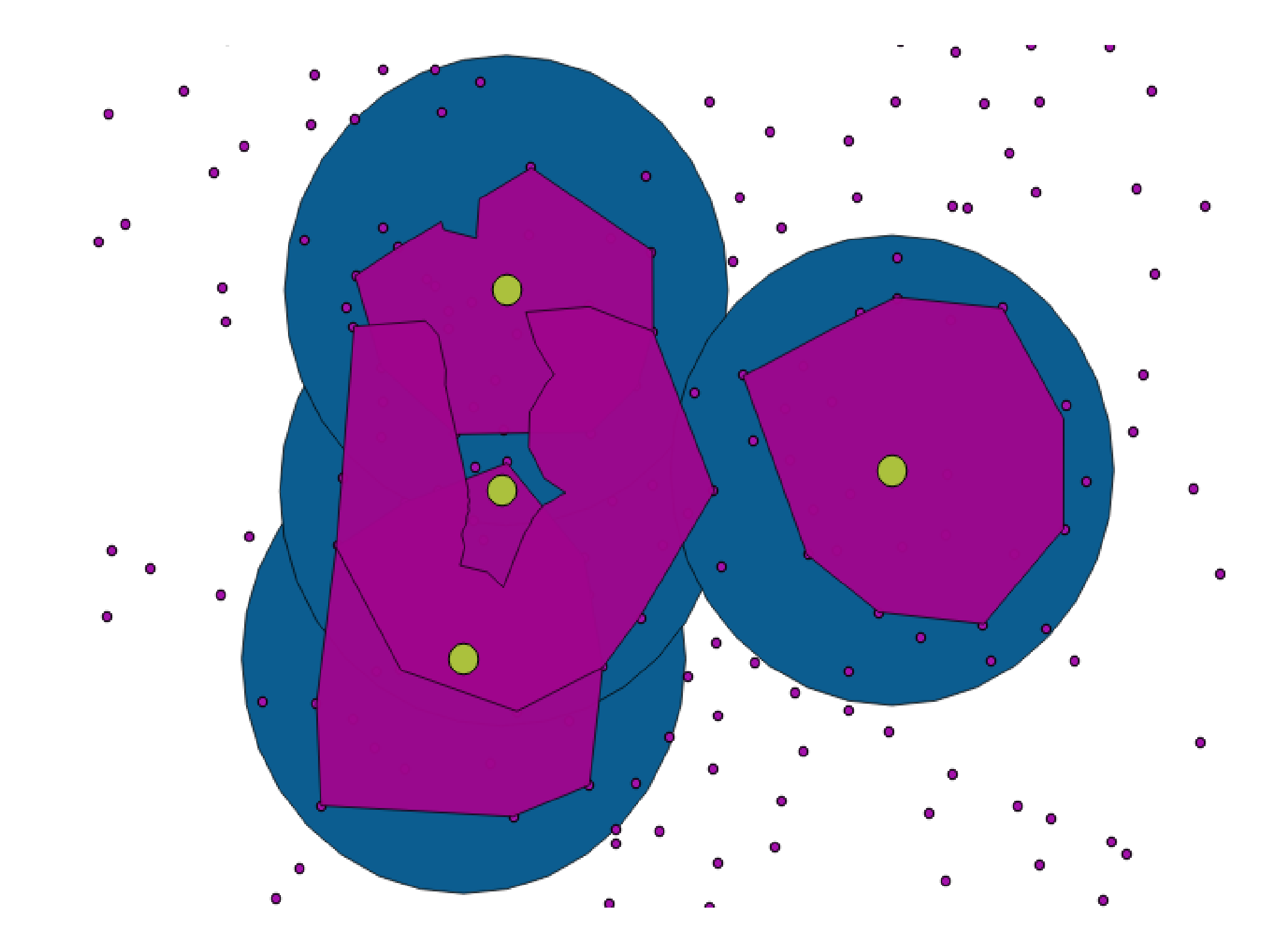


```
CREATE OR REPLACE FUNCTION signal_correlate(a float[], b float[])
RETURNS numeric
AS $$

from scipy import signal
import numpy as np

return np.argmax(signal.correlate(a, b)) - len(a)

$$ LANGUAGE plpythonu;
```

```
WITH
  a AS (
    SELECT *, bikes::numeric / (bikes + stands)::numeric AS avl, extract(hour FROM timestamp) h
    FROM gl.velov
    WHERE extract(isodow FROM timestamp) BETWEEN 1 AND 5 AND (bikes + stands) != 0
  )

  , b AS ( SELECT station id, h, avg(avl) AS avl FROM a GROUP BY id, h)

  , c1 AS ( SELECT id, h, avl FROM b WHERE id = 2012 ORDER BY h) -- Bellecour
  , c2 AS ( SELECT id, h, avl FROM b WHERE id = 3001 ORDER BY h) -- PartDieu
  , c3 AS ( SELECT id, h, avl FROM b WHERE id = 1002 ORDER BY h) -- Opera
  , c4 AS ( SELECT id, h, avl FROM b WHERE id = 2035 ORDER BY h) -- Republique


SELECT signal_correlate(array_agg(c1.avl), array_agg(c2.avl)) AS partdieu,
            signal_correlate(array_agg(c1.avl), array_agg(c3.avl)) AS opera,
            signal_correlate(array_agg(c1.avl), array_agg(c4.avl)) AS republique
 FROM c1, c2, c3, c4
 WHERE c1.h = c2.h AND c1.h = c3.h AND c1.h = c4.h
```

```
 partdieu | opera | republique
----------+-------+------------
        4 |    -1 |         -1
```

```
WITH t AS ( SELECT *, bikes::numeric / (bikes + stands)::numeric AS avl,
                        extract(hour FROM timestamp) h
              FROM gl.velov
              WHERE extract(isodow FROM timestamp) BETWEEN 1 AND 5
                  AND (bikes + stands) != 0 ),
    a AS ( SELECT station id, h, avg(avl) AS avl FROM t GROUP BY id, h ),
    h AS ( SELECT id, array_agg(avl) avl, array_agg(h) h FROM a GROUP BY id ),
    s AS ( SELECT id, ST_Transform(geom, 2154) AS geom FROM gl.station ),
    d AS ( SELECT s1.id s1, lat.id s2, lat.d d FROM s AS s1 ,
         LATERAL ( SELECT s2.id, ST_Distance(s1.geom, s2.geom) as d
                FROM s AS s2
                WHERE NOT s1.id > s2.id AND NOT ST_Equals(s1.geom, s2.geom)
                ORDER BY s1.geom <-> s2.geom LIMIT 25) AS lat
         WHERE lat.d < 1000 ),
    c AS ( SELECT s1, s2, d, signal_correlate(h1.avl, h2.avl) s
         FROM h h1, h h2, d WHERE d.s1 = h1.id AND d.s2 = h2.id ORDER BY s1, s2),
    g AS ( SELECT s1 id,array_agg(s2) AS ids FROM c WHERE s IN (-1, 0, 1) GROUP BY s1 ),
    z AS ( SELECT g.id, ST_ConcaveHull(ST_Collect(geom), 0.6) , row_number() OVER() i
        FROM g, s WHERE s.id = ANY (g.ids) GROUP BY g.id)


 SELECT * FROM z WHERE id = 2012 OR id = 3001 OR id = 1002 OR id = 2035
```



Skills to fully play with

SQL++

(Open) Data

PostGIS
ToolBox

Statistical
skills

PG Extension

Python

#Conclusion

PostgreSQL behaves like an extensible
and integrated Framework

(modern) SQL and Python acting as glue languages

Possible Bridge beetween
GIS and Python DataScience communities

Could be that fun !

Thanks

http://www.oslandia.com

https://github.com/Oslandia/presentations