Spatial Analysis in CyberGIS

towards a spatial econometrics workbench

Luc Anselin, Sergio Rey and Myunghwa Hwang
GeoDa Center
School of Geographical Sciences and Urban Planning
Arizona State University

- Acknowledgment

  - NSF Award OCI-1047916: CyberGIS Software Integration for Sustained Geospatial Innovation

- Team
  - Luc Anselin
  - Sergio J. Rey
  - Robert Pahle
  - Myunghwa Hwang
  - Philip Stephens
  - Xing Kang

Goals

towards cyberinfrastructure in support of spatial analysis (Goodchild, 2010)

GeoDa Center for Geospatial Analysis and Computation
Arizona State University

- Some Specific Project Goals
  - Integrate and sustain a core set of composable, interoperable, manageable, and reusable CyberGIS software elements based on community-driven and open source strategies

- Some Specific Project Goals (2)
  - Empower high-performance and scalable CyberGIS by exploiting spatial characteristics of data and analytical operations for achieving unprecedented capabilities for geospatial knowledge discovery

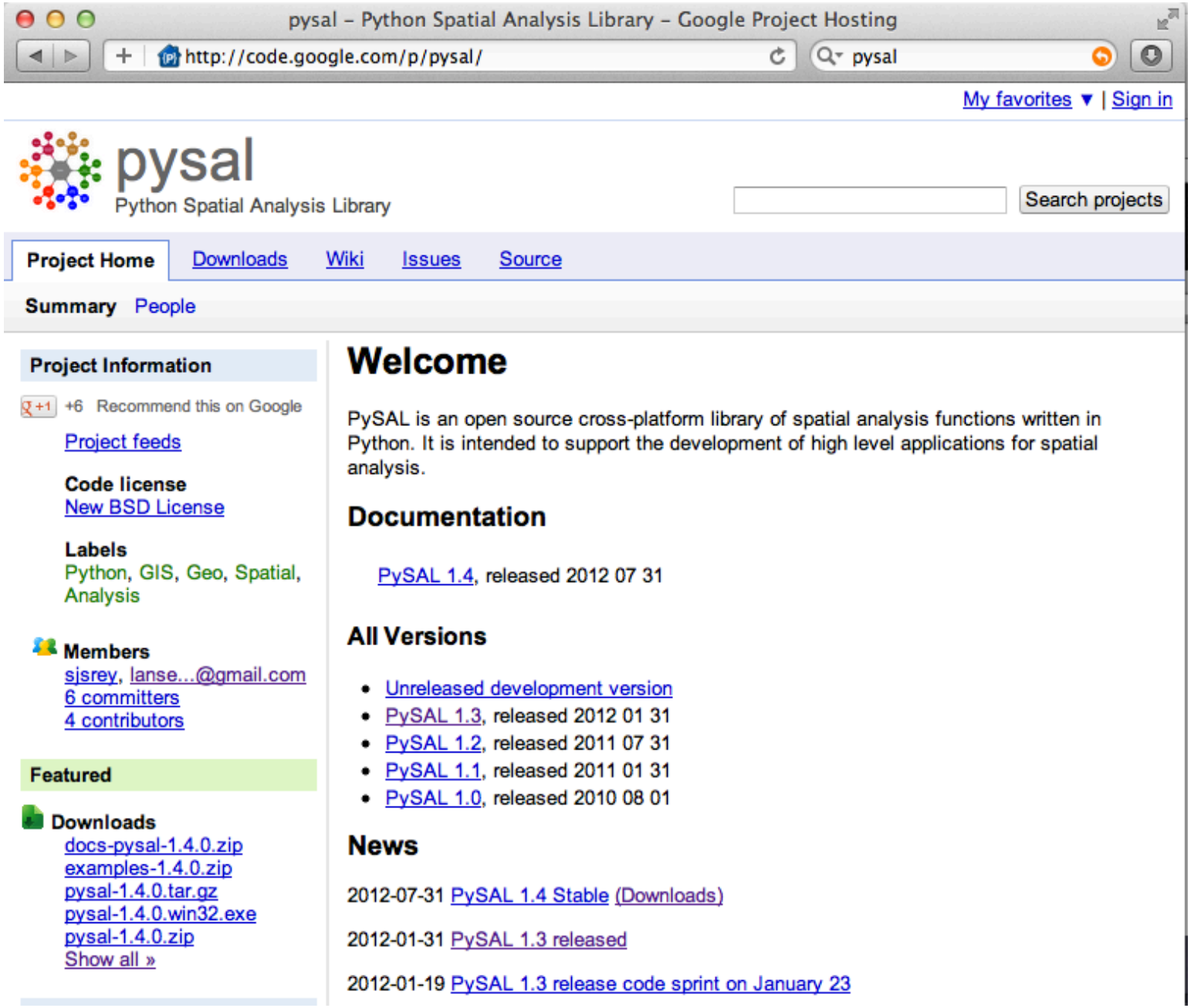
- Challenge
  - most current spatial analysis/spatial econometrics software written for single CPU
  - rethink and rewrite analytical, algorithmic and processing facilities to integrate into a cyberinfrastructure
  - lack of interoperability

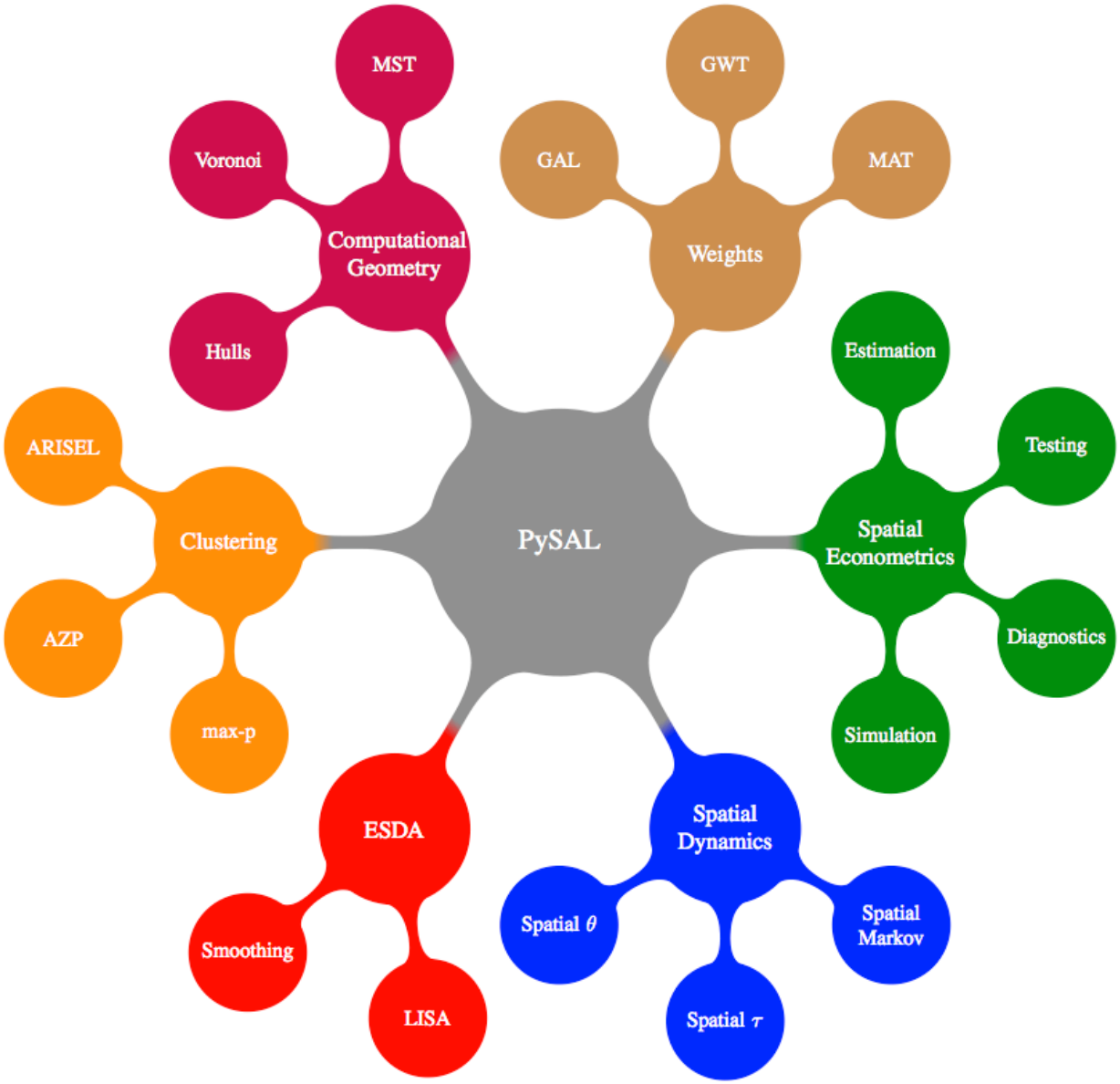
- Spatial Econometrics Workbench
  - framework for supporting spatial econometric research in a cyberscience era (Anselin and Rey, IJGIS 2012)
  - support for scientific workflow
  - leverages PySAL

Leveraging PySAL

GeoDa Center
for Geospatial Analysis
and Computation
Arizona State University

- PySAL
  - open source library of Python routines for spatial analysis: geocomputation, spatial weights, spatial autocorrelation, spatial econometrics
  - http://pysal.org
  - hosted on google code





GeoDa Center
for Geospatial Analysis
and Computation

ARIZONA STATE UNIVERSITY

- PySAL Functions for CyberGIS
  (selected examples)
  - spatial weights
  - global/local spatial autocorrelation
  - simulation estimators

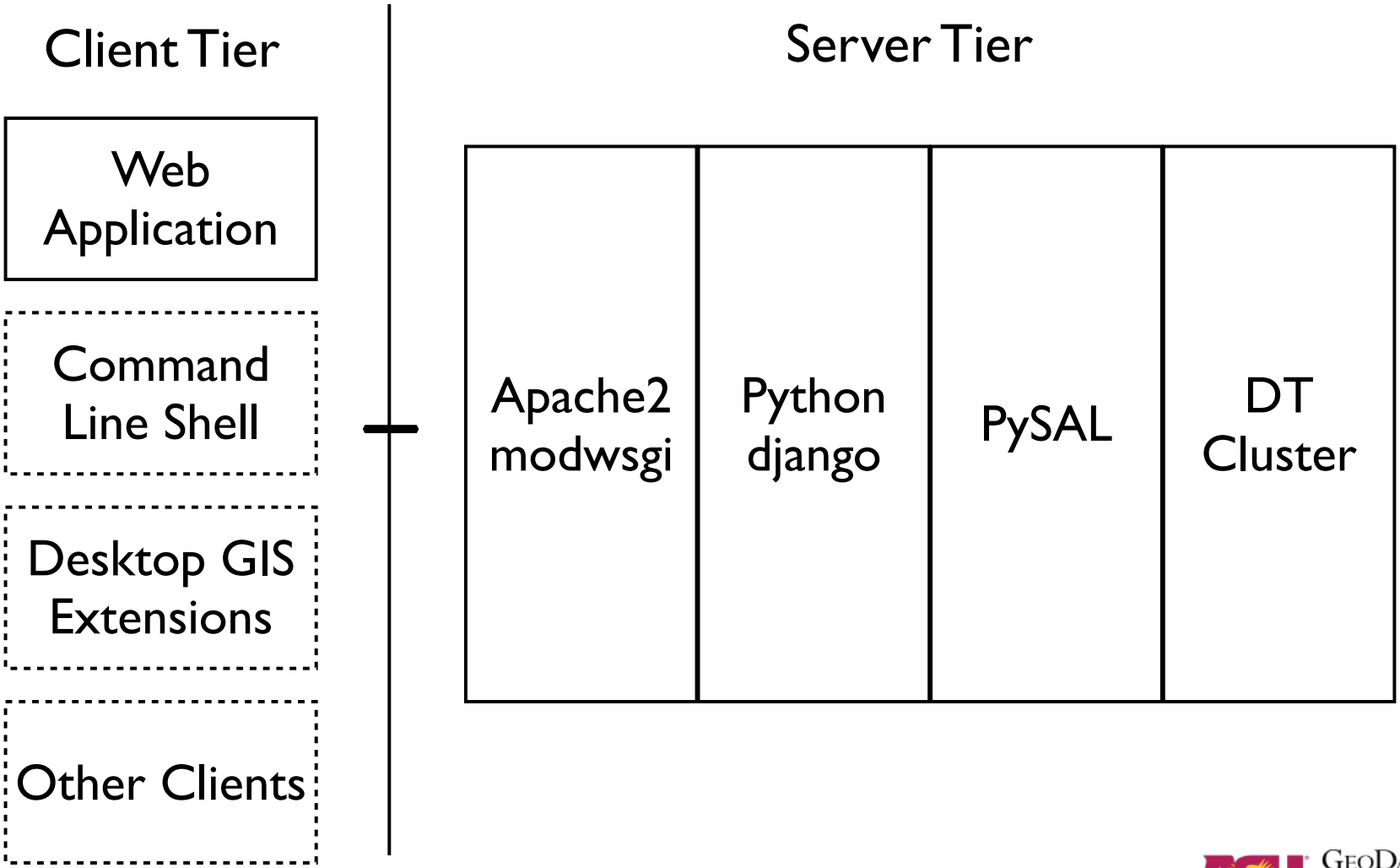
- Migrating to CyberGIS
  - provide functionality as web services
  - performance = need for parallelization + refined algorithms
  - interoperability = need for metadata and provenance tracking

Web Services

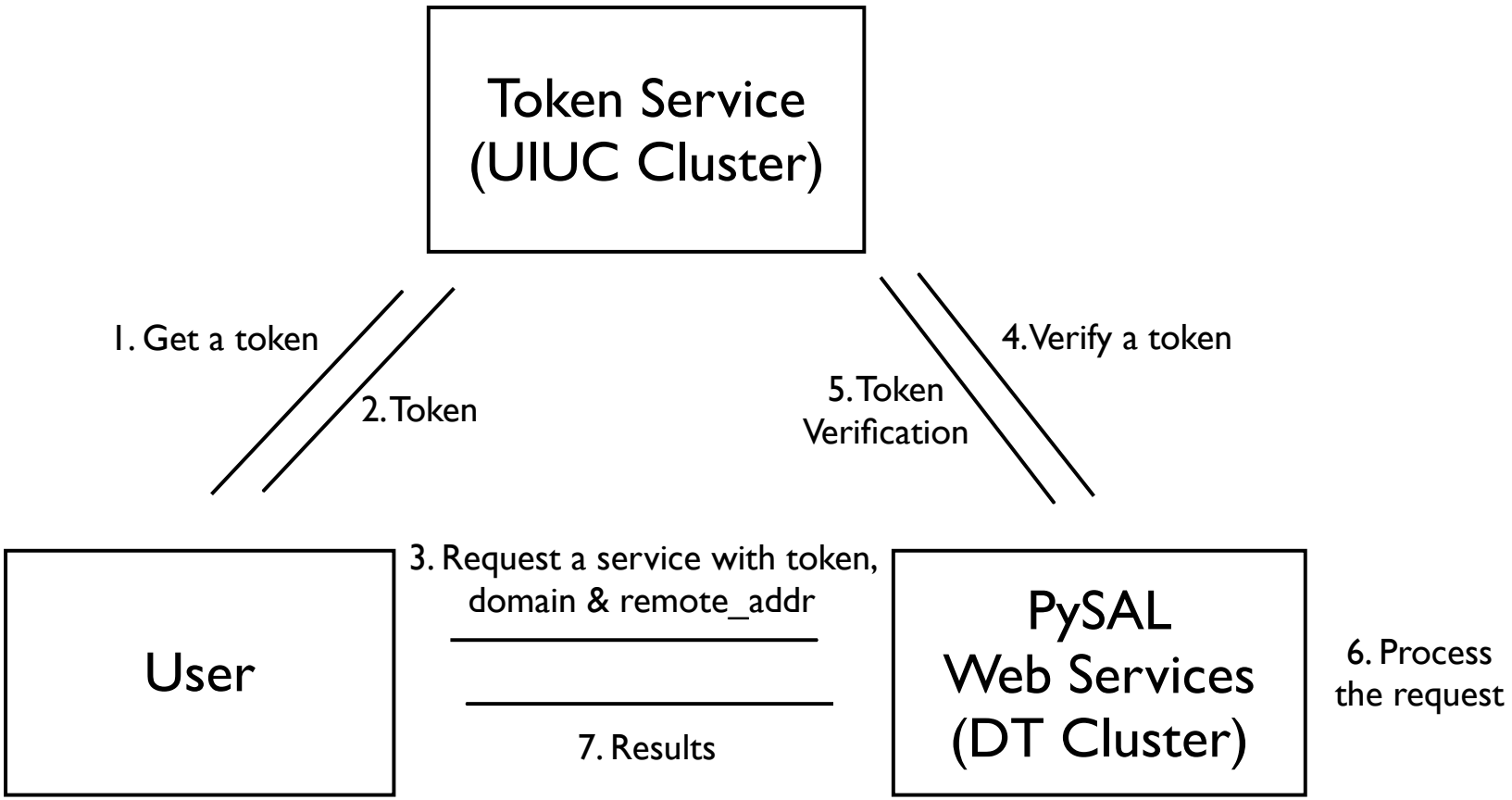
GeoDa Center
for Geospatial Analysis
and Computation
Arizona State University

- PySAL Web Services
  - wrapping PySAL code as services
  - providing access through portal
  - infrastructure = link between UIUC and ASU Decision Theater Cluster
  - prototype: spatial weights operations

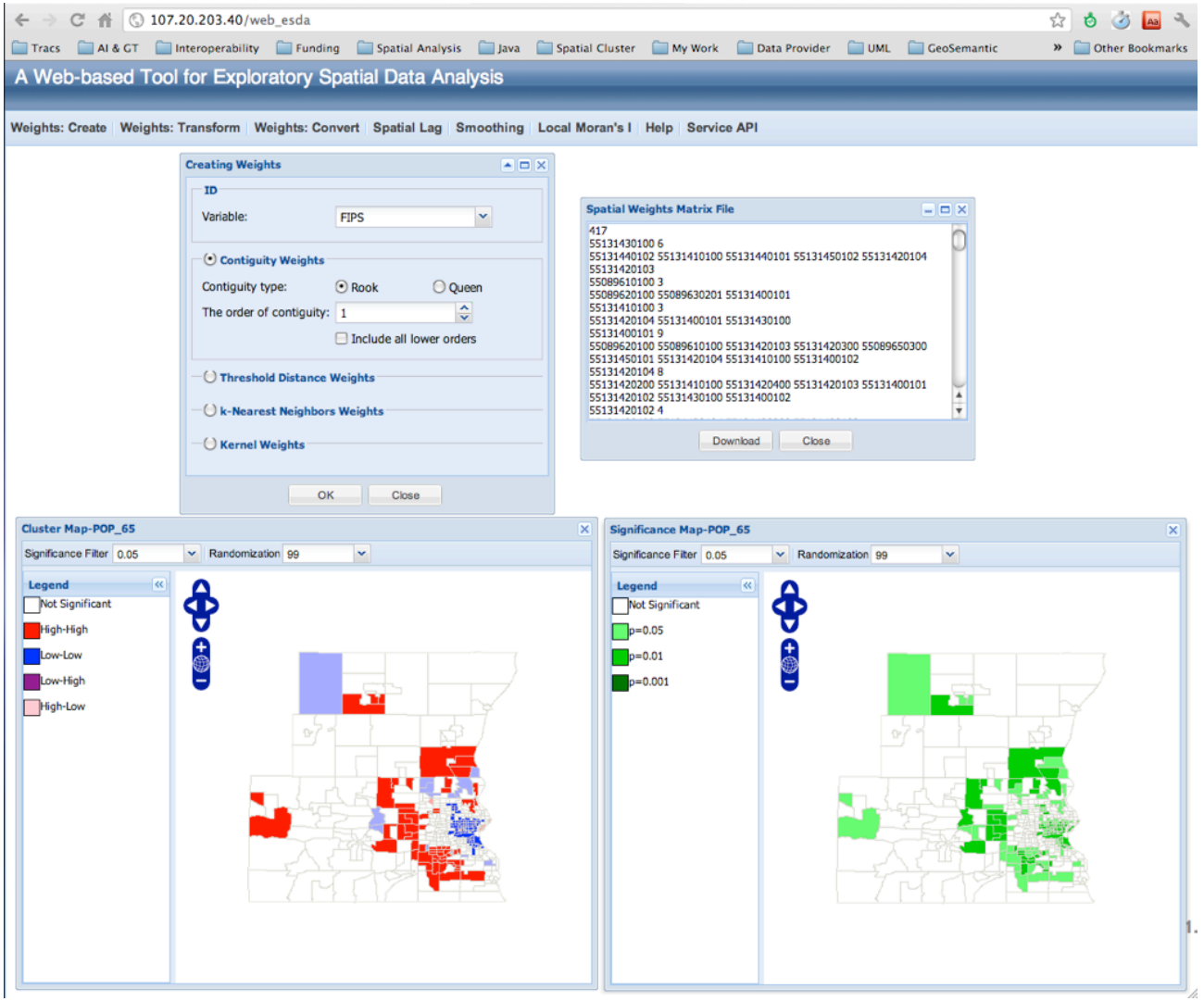
Service Implementation



Integration with Token Service



Web 2.0 Interface to PySAL Services



GeoDa Center for Geospatial Analysis and Computation
Arizona State University

- Command-Line Access
  - open to cyberGIS gateway users
  - authentication: gateway token
  - Examples:

```
./kernelW.py -g ohio.zip -d 2 -f false -k 5 -t quadratic -i FIPSNO
./knnW.py -g ohio.zip -d 2 -k 5 -i FIPSNO
./lag.py -d ohio.dbf -y POPMW68 -w queen.gal
./moran.py -d ohio.dbf -y POPMW68 -w queen.gal -m global -s 99
```

Parallelization

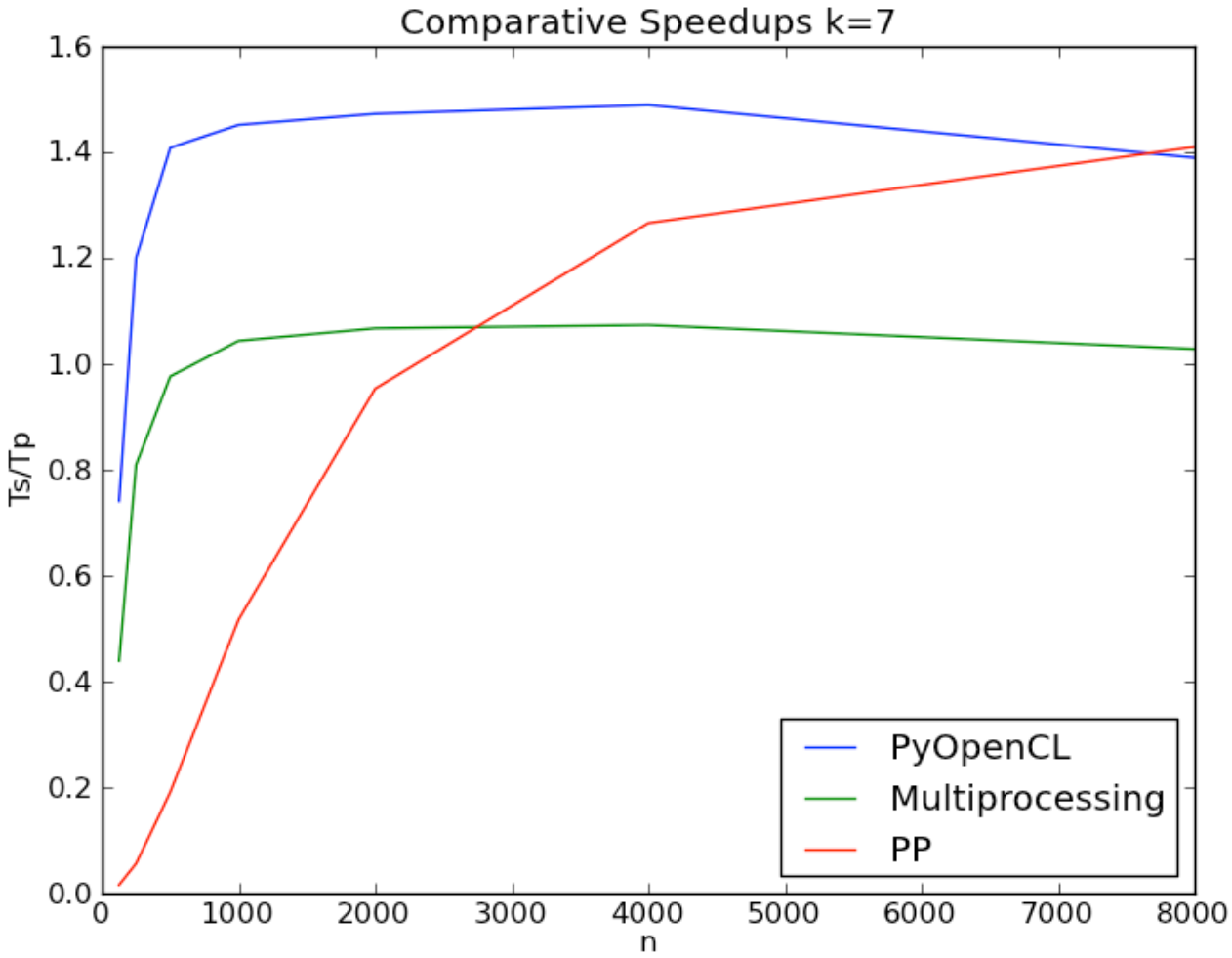
- Why?
  - real-time analysis in a decision support environment
  - 8 second rule
  - spatial econometric methods: spatial weights manipulations, simulation estimators

- Approach
  - focused on PySAL and use of Python parallelization modules
  - multicore and GPU, clusters (ongoing)
  - available Python libraries:
    pyopencl, multiprocessing, parallel python
  - not all spatial analysis methods suitable for parallelization

- Initial Results
  - embarrassingly parallel:
    permutation inference (simulations), local statistics, simulation estimators
  - not so straightforward:
    sequential analysis combined with parallel
  - experimentation needed:
    results not always as expected
    no obvious general results

- Some Examples
  - implementing parallel Lisa (Local Moran)
  - Fisher-Jencks map classification
  - regionalization (contiguity-constrained clustering)

Example: Fisher-Jencks Map Classification



Provenance and Standards

GeoDa Center
for Geospatial Analysis
and Computation
Arizona State University

- Scientific Workflow
  - data provenance
  - model provenance
  - lineage of data sources, manipulation and analytical operations (DAG)
  - enable replication

- Specialized Needs
  - spatial operations combine data with processing
  - requires metadata for attribute data
  - requires metadata for operations

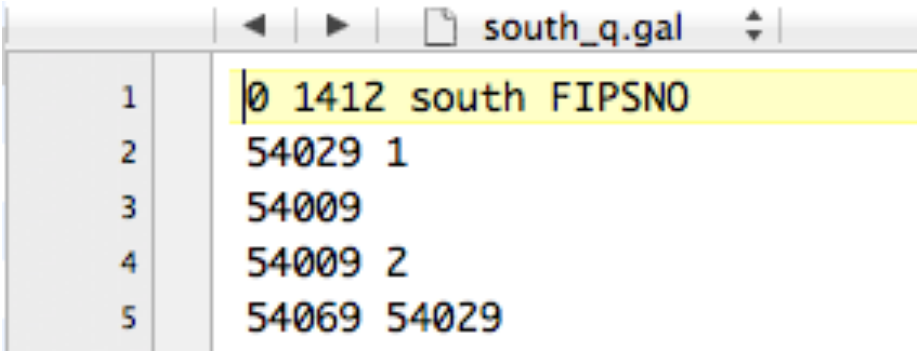
- Example: Spatial Weights
  - provenance = lineage
  - includes spatial data source, type of weights (e.g., contiguity, distance), any standardization or manipulation (e.g., higher order)

```
----------
SUMMARY OF OUTPUT: SPATIAL TWO STAGE LEAST SQUARES ESTIMATION
-----------------------------------------------------------
Data set            :    south.dbf
Weights matrix      : File: south_k6.gwt
Dependent Variable  :        HR60  Number of Observations:        1412
Mean dependent var  :      7.2921  Number of Variables   :           6
S.D. dependent var  :      6.4210  Degrees of Freedom    :        1406

Pseudo R-squared    :    0.138208
Spatial Pseudo R-squared:    0.068132

HAC Standard Errors; Kernel Weights: File: south_ep_k20.kwt
------------------------------------------------------------------------------
    Variable     Coefficient       Std.Error     z-Statistic     Probability
------------------------------------------------------------------------------
    CONSTANT       1.3694706       1.5637818       0.8757428       0.3811699
        RD60       1.0183869       0.4345945       2.3433038      0.01911381
        PS60       0.1348746       0.4603548       0.2929797       0.7695376
        UE60      -0.0676808       0.0933944      -0.7246778       0.4686497
        DV60       0.4764954       0.1956288       2.4357123       0.0148625
      W_HR60       0.6482585       0.1875803       3.4558989    0.0005484609
------------------------------------------------------------------------------
Instruments: W_RD60, W_PS60, W_UE60, W_DV60
```

Example: PySAL spgreg

what do we know about south_k6.gwt and south_ep_k20.kwt

- Early Attempts
  - header line in gal and gwt files
  - "standard" initially in SpaceStat and later adopted by R, GeoDa, PySAL, etc.
  - insufficient



- Taxonomy of Weights Formats in PySAL

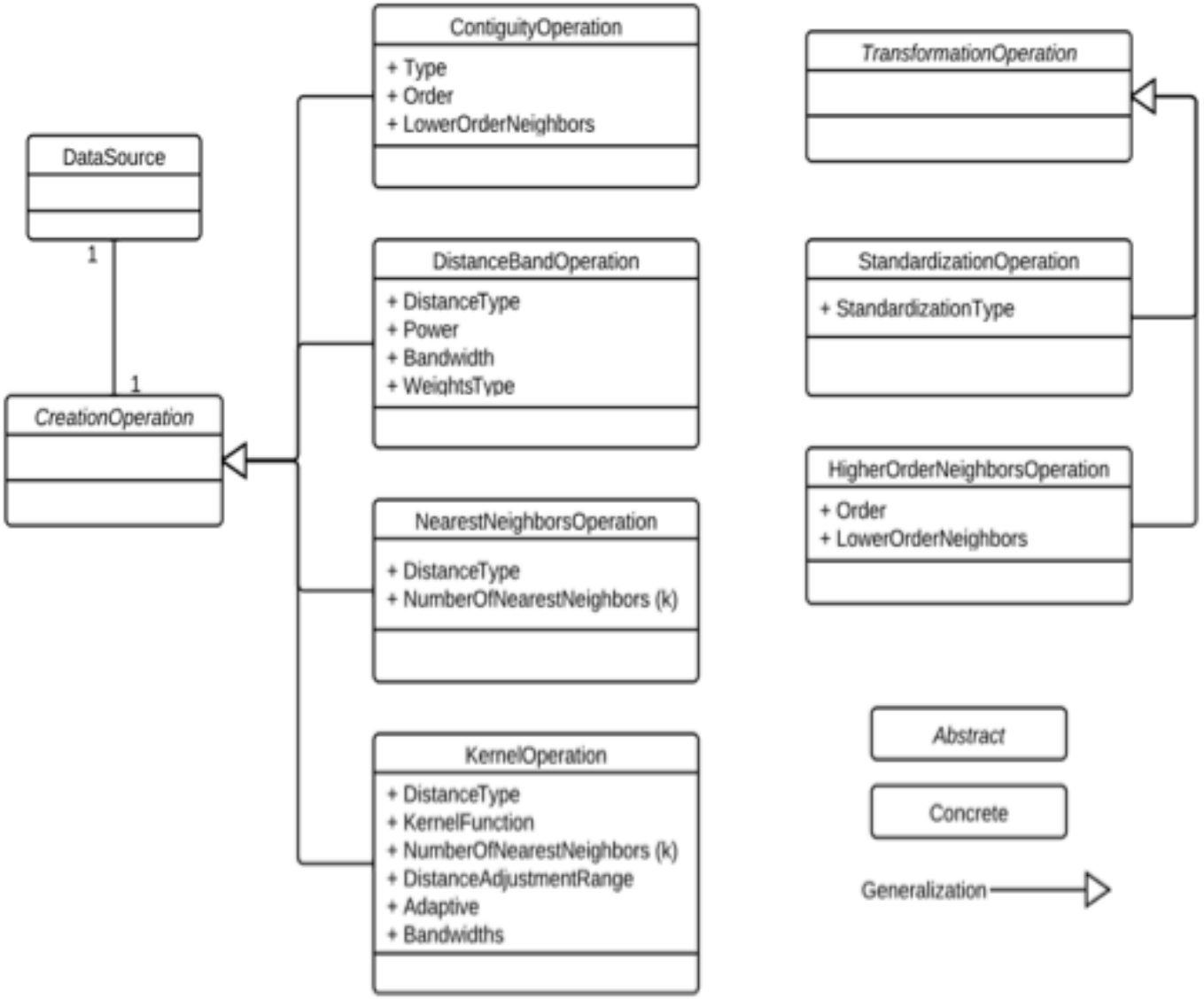
```
>>> pysal.open.check()
PySAL File I/O understands the following file extensions:
Ext: '.shp', Modes: ['r', 'wb', 'w', 'rb']
Ext: '.mtx', Modes: ['r', 'w']
Ext: '.swm', Modes: ['r', 'w']
Ext: '.mat', Modes: ['r', 'w']
Ext: '.shx', Modes: ['r', 'wb', 'w', 'rb']
Ext: '.stata_text', Modes: ['r', 'w']
Ext: '.geoda_txt', Modes: ['r']
Ext: '.dbf', Modes: ['r', 'w']
Ext: '.dat', Modes: ['r', 'w']
Ext: '.gwt', Modes: ['r', 'w']
Ext: '.gal', Modes: ['r', 'w']
Ext: '.arcgis_text', Modes: ['r', 'w']
Ext: '.wk1', Modes: ['r', 'w']
Ext: '.arcgis_dbf', Modes: ['r', 'w']
Ext: '.geobugs_text', Modes: ['r', 'w']
Ext: '.csv', Modes: ['r']
Ext: '.wkt', Modes: ['r']
```

- Conceptual Framework
  - separate data source from operations
  - data source: polygon or coordinate files with standard metadata (projection, origin, etc.)
  - operations: no metadata yet (initial work in OGC WPS etc. but not for analysis)

- Operations
  - creating the network structure
    0-1 connectivity graph
  - computing the weights values

- Prototype Weights Metadata
  - content model = vocabularies to describe spatial weights provenance and operations
  - representation = XML Schema

Core Operations



Extract of XML Schema

```
<xs:complexType name="CreationOperationType" abstract="true">
    <xs:complexContent>
        <xs:extension base="wgt:WeightsOperationType">
            <xs:sequence>
                <xs:element ref="wgt:DataSource"/>
            </xs:sequence>
        </xs:extension>
    </xs:complexContent>
</xs:complexType>
<xs:complexType name="ContiguityOperationType">
    <xs:complexContent>
        <xs:extension base="wgt:CreationOperationType">
            <xs:sequence>
                <xs:element name="Type" type="xs:string"/>
                <xs:element name="Order" type="xs:int"/>
                <xs:element name="LowerOrderNeighbors" type="xs:boolean"/>
            </xs:sequence>
        </xs:extension>
    </xs:complexContent>
</xs:complexType>
```

```
<?xml version="1.0" ?>
<weights:WeightsMetadata xmlns:operations="http://www.swm.org/2011/operations"
    xmlns:weights="http://www.swm.org/2011/weights"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
    <weights:WeightsMatrix>
        <weights:WeightsOperation xsi:type="operations:ContiguityOperationType">
            <weights:TimeStamp>2012-01-18T00:46:11</weights:TimeStamp>
            <weights:Software>
                <weights:Name>GeoDa Center Weights Service</weights:Name>
                <weights:Version>0.1</weights:Version>
            </weights:Software>
            <weights:DataSource>
                <weights:TemporalCoverage>
                    <weights:StartTime>2010-01-01T12:00:00</weights:StartTime>
                    <weights:EndTime>2010-07-01T12:00:00</weights:EndTime>
                </weights:TemporalCoverage>
                <weights:SpatialCoverage>
                    <weights:PlaceName>United States,U.S.,State or Equivalent
                        Entity,Arizona,AZ,04</weights:PlaceName>
                    <weights:GeographicExtent>
                        <weights:MinX>-114.816591</weights:MinX>
                        <weights:MinY>31.332177</weights:MinY>
                        <weights:MaxX>-109.045223</weights:MaxX>
                        <weights:MaxY>37.00426</weights:MaxY>
                    </weights:GeographicExtent>
                    <weights:GeometryType>Polygon</weights:GeometryType>
                    <weights:GeographicCoordinateSystem>
                     .
                     </weights:GeographicCoordinateSystem>
                    <weights:Scale>Census Tract,Tract</weights:Scale>
                </weights:SpatialCoverage>
<weights:DataLocation>/ [Path]/arizona_tracts_2010/tl_2010_04_tract10.shp</weights:DataLocation>
                <weights:DataType>ESRI Shape File</weights:DataType>
                <weights:IdVariable>GEOID10</weights:IdVariable>
                <weights:NumberOfObservations>1526</weights:NumberOfObservations>
            </weights:DataSource>
            <operations:Type>Rook</operations:Type>
            <operations:Order>1</operations:Order>
            <operations:LowerOrderNeighbors>true</operations:LowerOrderNeighbors>
        </weights:WeightsOperation>
    </weights:WeightsMatrix>
</weights:WeightsMetadata>
```

- Moving Forward
  - automatic generation of provenance information
  - automatic retrieval
  - semantic interoperability

Roadmap

- Methodological Refinements
  - parallelizing spatial methods and algorithm improvement
  - spatial probit/tobit
  - space-time methods (space-time scan statistic)

- Operationalization
  - web services / portal
  - weights metadata
  - automatic provenance generation
  - spatial econometrics workbench

- Community Building
  - demonstrate usefulness
  - demonstrate added value
  - training and documentation
  - facilitate participation