

Data Science in ArcGIS Using Python and R

Kade Beem
Lu Zhang

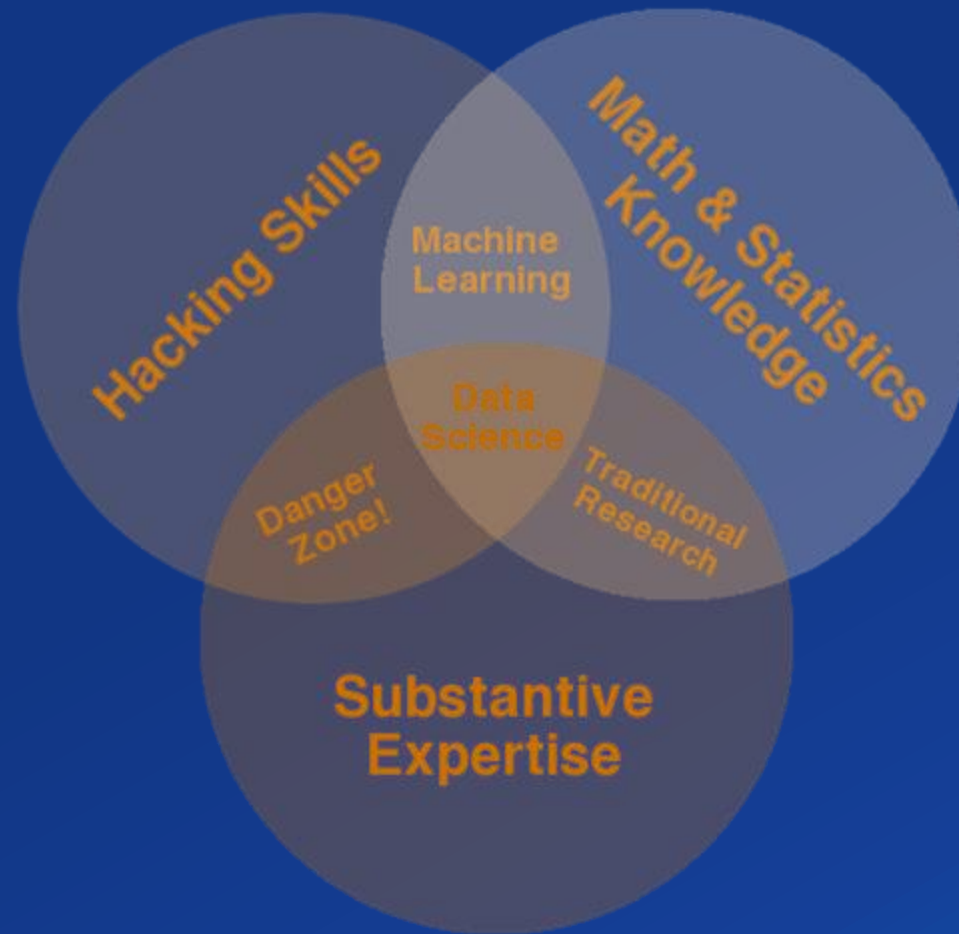
2020 ESRI FEDERAL GIS CONFERENCE | WASHINGTON, D.C.

Agenda

- Introduction
- Performing Analysis using Python
 - ArcPy
 - ArcGIS API for Python
 - Demonstration
- Utilizing the R-ArcGIS Bridge
 - Installation / Use Cases
 - Sample Workflow
- Q&A

Data Science

Introduction



Drew Conway 2010

Data Science

From Core to Community

- Core analytics in ArcGIS
 - Maximize performance and utility
 - E.g. Spatial Statistics, Geostatistics, Spatial Analyst
 - E.g. GeoAnalytics, Insights, ArcGIS Python SDK
- The interoperability of the ArcGIS platform makes workflows more efficient
 - Techniques and methodologies continue to develop
 - Data availability continues to increase
- The data science community is vast and evolving
 - ArcGIS extends directly via scripting APIs
 - e.g. **Python**, **R**, Java



Spatial Analysis and Data Science Framework

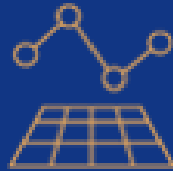
Patterns of Use



Data
Engineering



Visualization
& Exploration



Spatial
Analysis



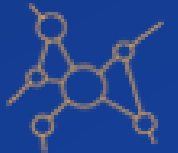
Machine
Learning & AI



Big Data
Analytics



Modeling
& Scripting



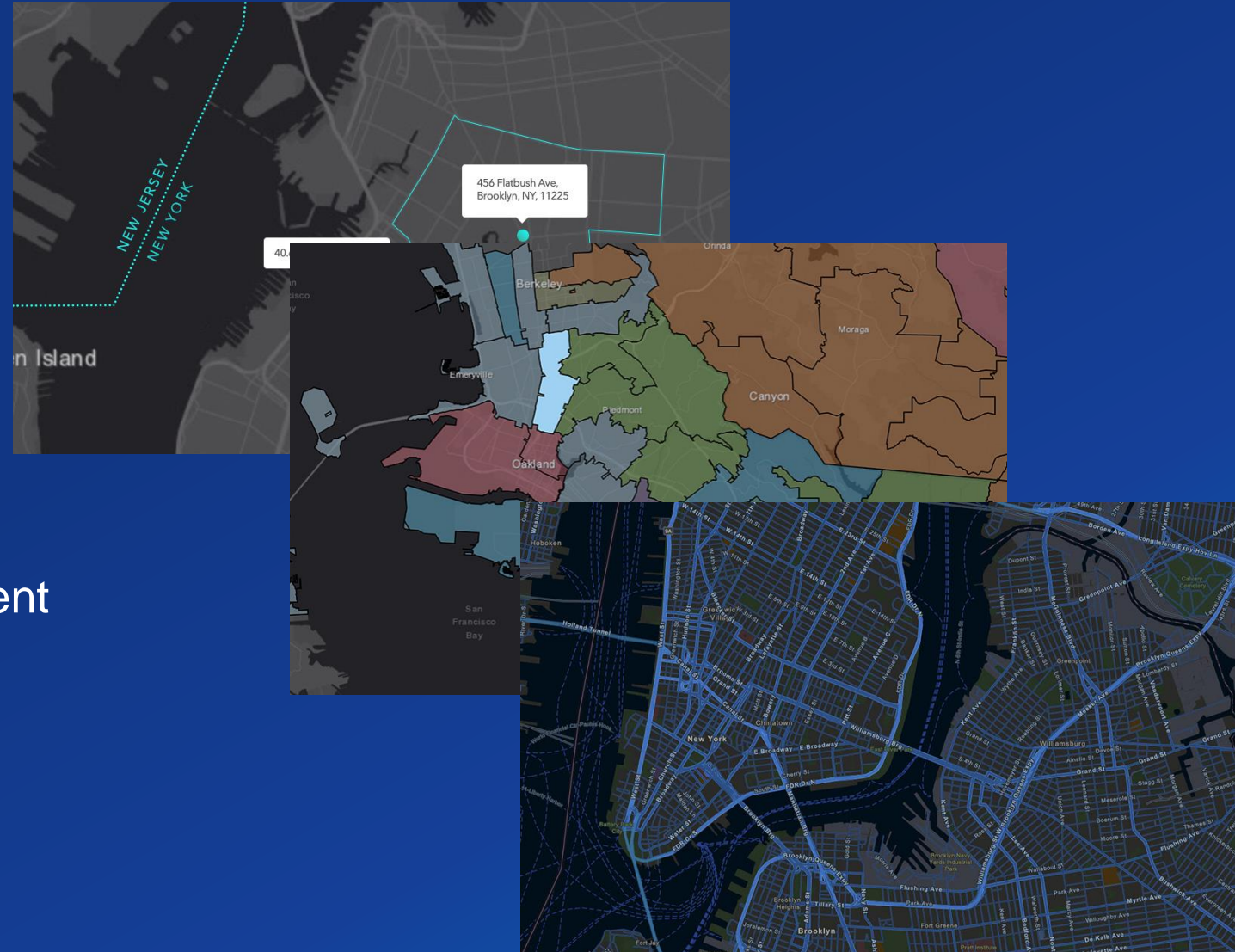
Sharing
& Collaboration



Spatial Analysis and Data Science Framework

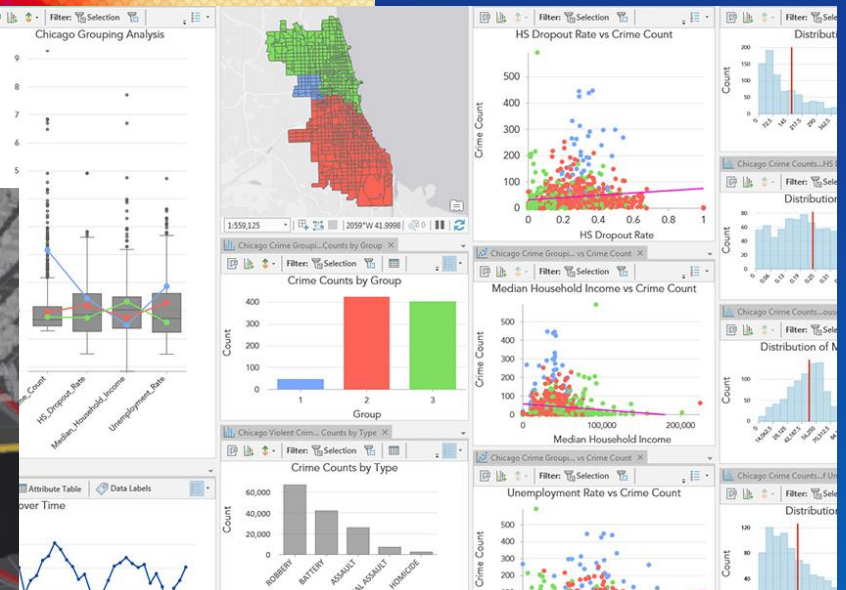
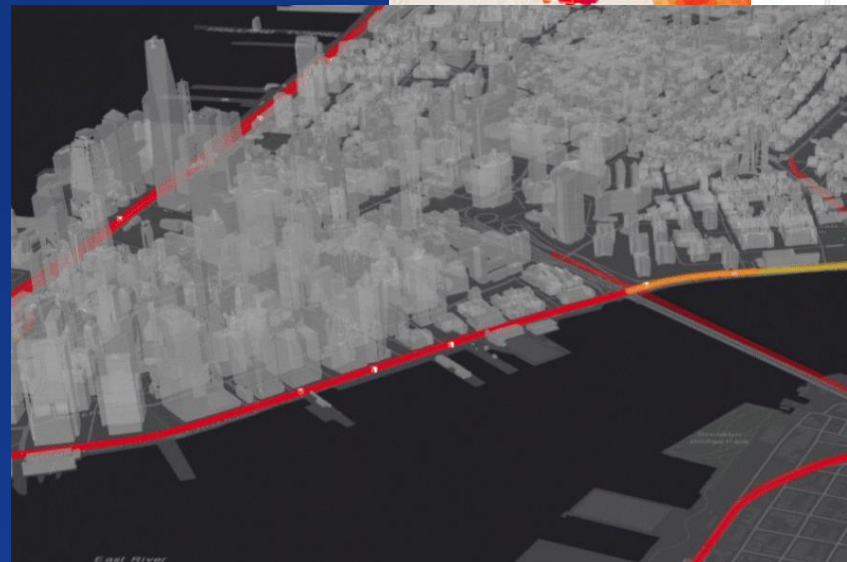
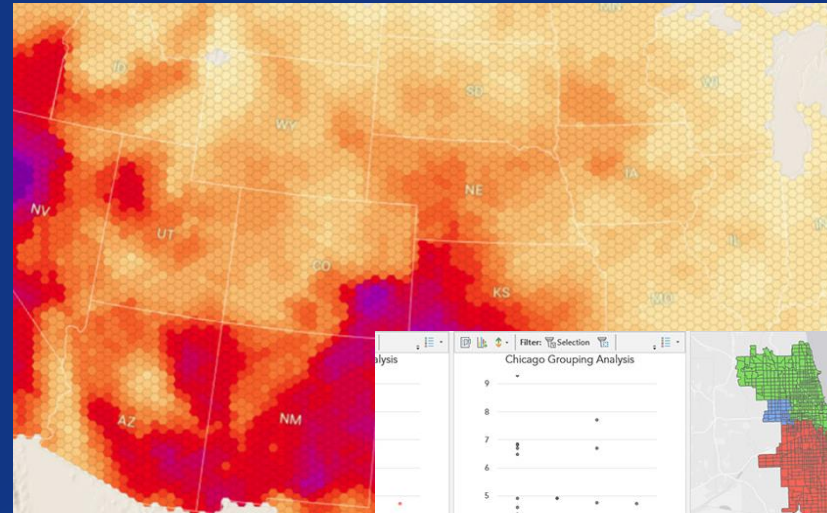
Data Engineering

- Extract and prep
- Location-enable your data
- Enrich your data
- Access ready-to-use data and content



Visualization & Exploration

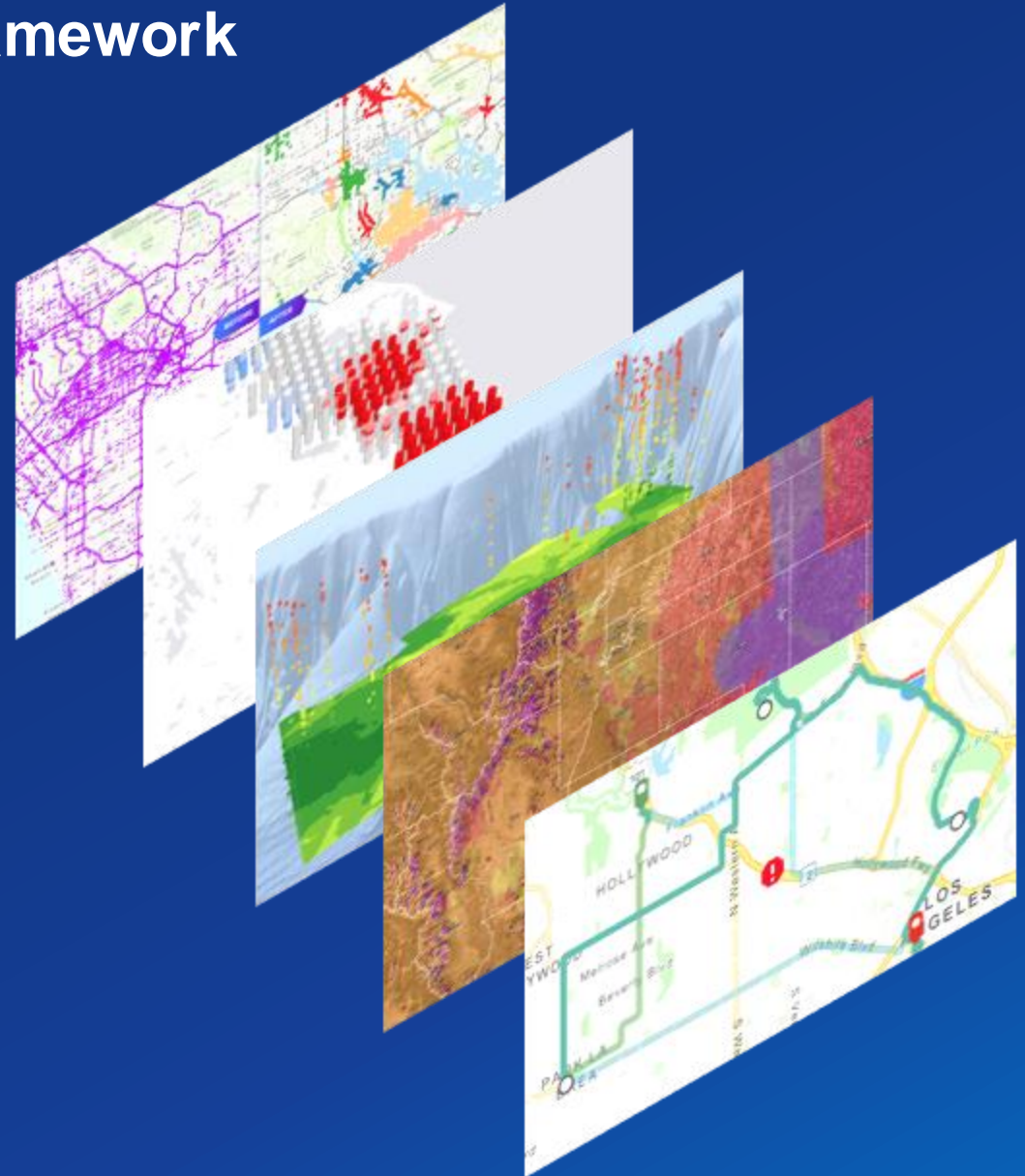
- Mapping
- Charting
- Animation



Spatial Analysis and Data Science Framework

Spatial Analysis

- Pattern detection
- Space-time analytics
- Predictive analytics
- Suitability modeling and location optimization
- Route planning and logistics
- Terrain analysis
- Scale up your analysis

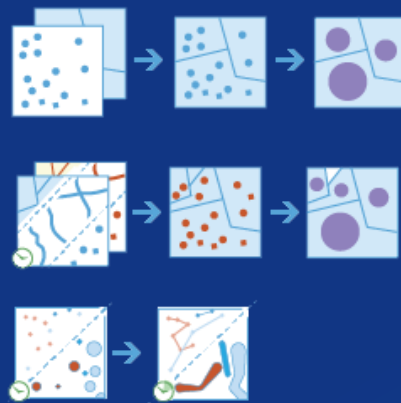


Spatial Analysis and Data Science Framework

Machine Learning & Artificial Intelligence

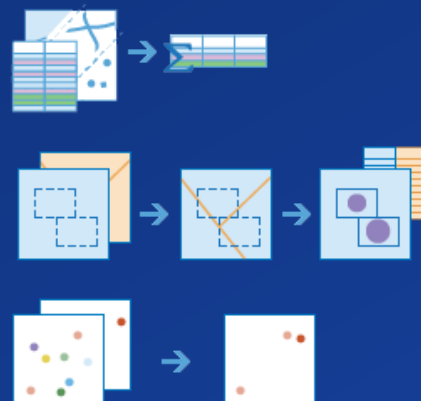
Classification

- Maximum Likelihood Classification
- Random Trees
- Support Vector Machine
- Find Dwell Locations



Clustering

- Spatially Constrained Multivariate Clustering
- Multivariate Clustering
- Density-based Clustering
- Find Point Clusters
- Hot Spot Analysis
- Cluster and Outlier Analysis
- Space Time Pattern Mining



Prediction

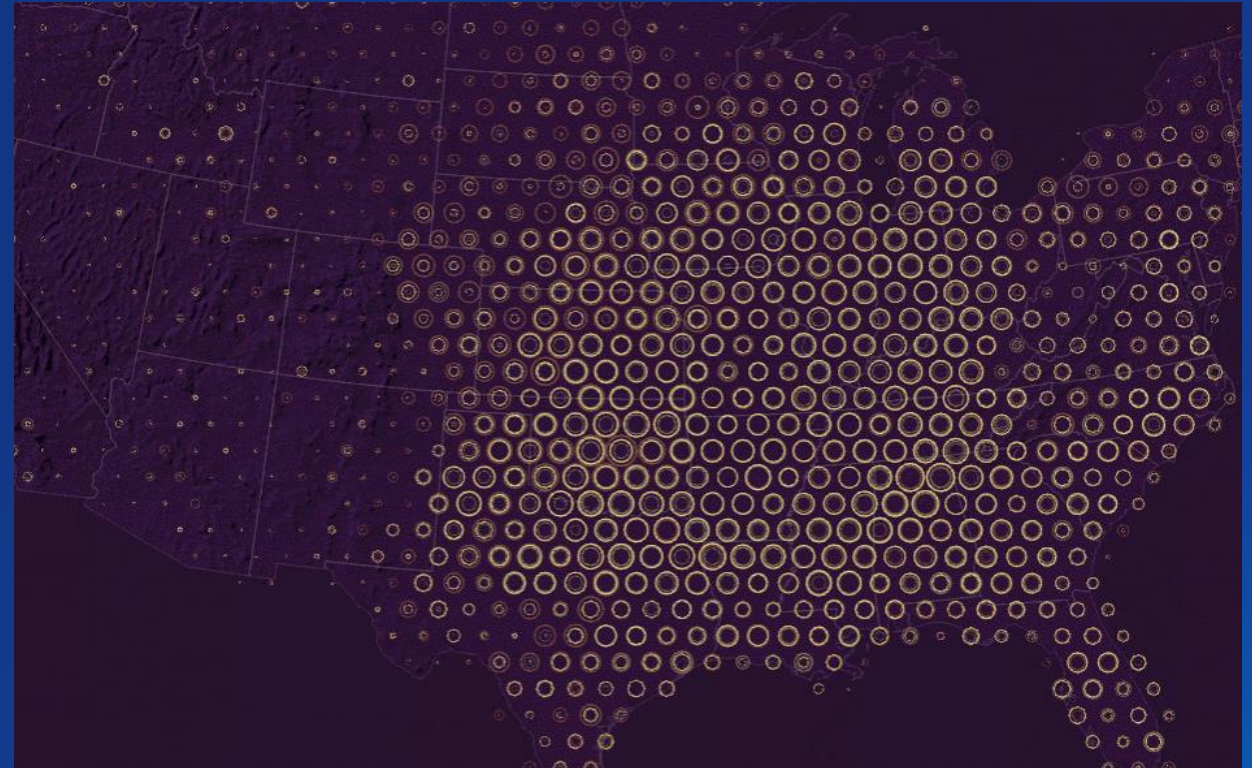
- Empirical Bayesian Kriging
- Areal Interpolation
- EBK Regression Prediction
- Ordinary Least Squares Regression and Exploratory Regression
- Geographically Weighted Regression
- Forest-based Classification and Regression



Spatial Analysis and Data Science Framework

Big Data Analytics

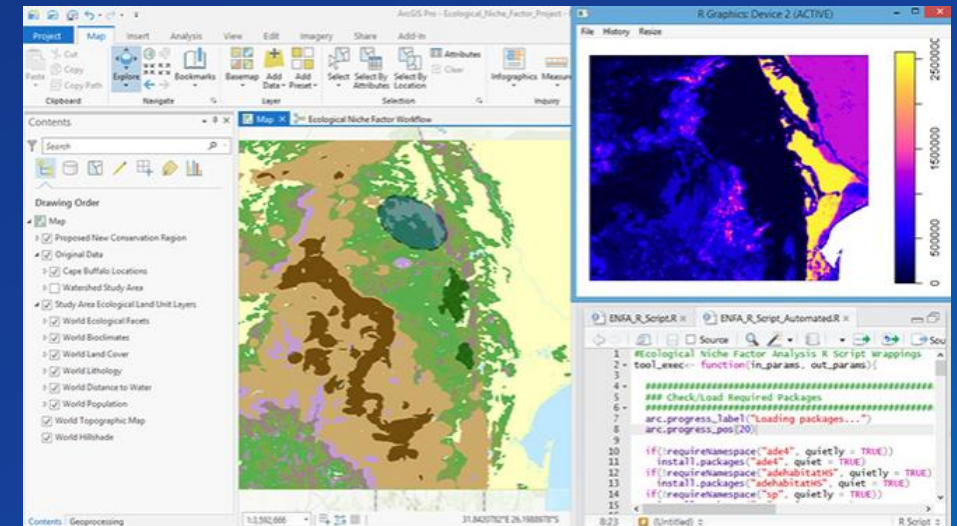
- **Visualization**
- **Distributed computing**
 - GeoAnalytics Server
 - GeoAnalytics Desktop
- **Imagery and raster analytics**
- **Real-time analytics**
 - GeoEvent Server



Spatial Analysis and Data Science Framework

Modeling & Scripting

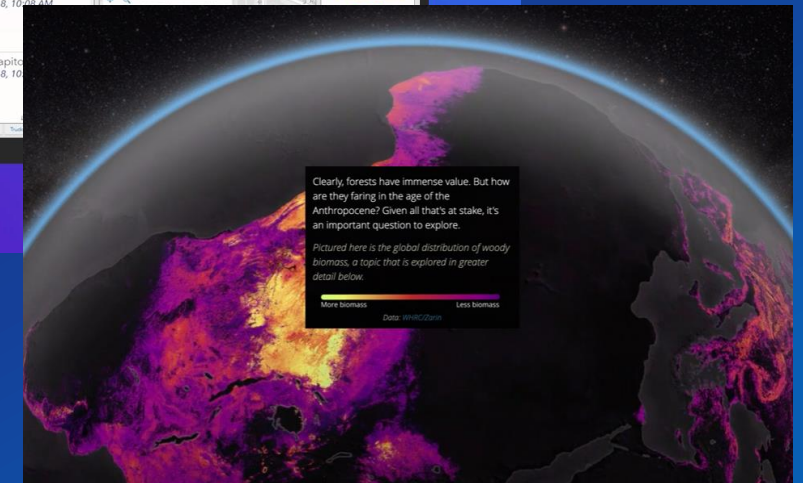
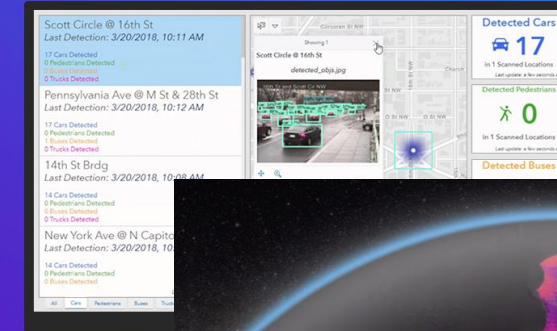
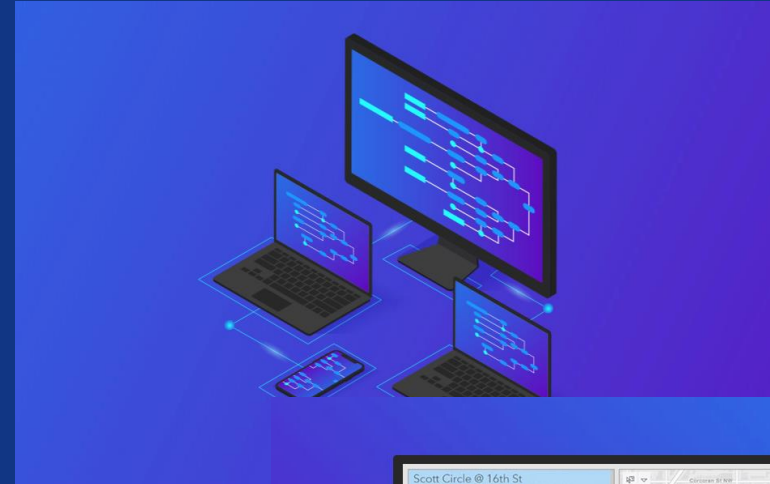
- Train sophisticated models
- Save time by automating
- Extensibility and integration
- Transparency and reproducibility



Spatial Analysis and Data Science Framework

Sharing & Collaboration

- **Work collaboratively and share result**
- **Integrate analytics across your organization**
- **Tell impactful stories to an engage audience**



Data Science Community

Python

IP[y]: IPython
Interactive Computing



arcgis.learn



Data Science Community

R

- Well over 12,000 packages to enhance core
- Most widely used statistical software in the world
- Diverse and powerful
 - Universities, Government, Industry
 - Finance, Ecology, Statistics
 - Machine learning, predictive analytics



Battle of Bands

Which one is best?



General Programming Language

Functionality

Tailored towards statistics and data analysis

Yes (PyPI and Conda)

Package Management

Yes (CRAN and Conda)

Individual machines to distributed computing environments

Scalability

Standalone computer or individual servers

Large number of libraries for graphical display of data

Display of Data

Numerous libraries for making incredible graphics

General Programming, Data Science, Web Development

Use Cases

Lingua franca of data science

YES!

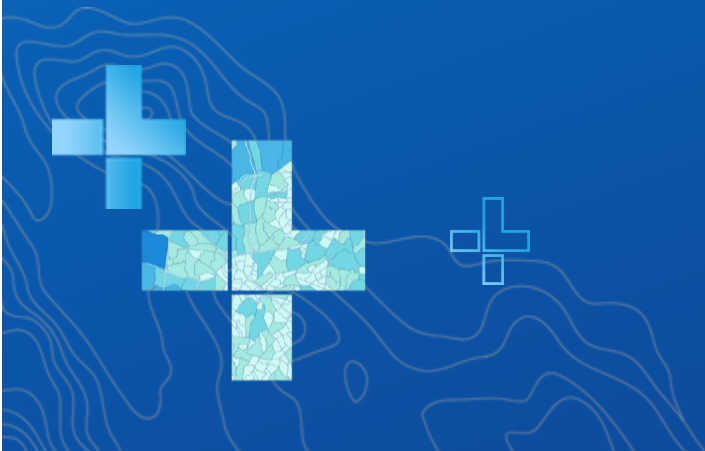
Integrates with ArcGIS

YES!

What are you most comfortable with?

What is the best tool for the job?

Performing Analysis Using Python



Python

ArcPy vs ArcGIS API for Python

ArcPy

- Use, extend, automate, desktop GIS
- Access GP tools, extensions, functions, and classes
- Create standalone scripts to run with an IDE
- Create script tools and added to a toolbox within ArcGIS Pro

ArcGIS API for Python

- Use, extend, automate, Web GIS
- Perform visualization, analysis, spatial data management, and GIS system administration
- Run interactively in a notebook
- Create script tools and added to a toolbox within ArcGIS Pro

Python

ArcPy

- Tools
- Functions
- Classes
- Modules
 - Mapping – arcpy.mp
 - Manipulate contents of ArcGIS Pro projects and layer files
 - Upgraded from arcpy.mapping
 - Leverages Python 3.x

```
import arcpy
aprx = arcpy.mp.ArcGISProject(r"C:\Projects\YosemiteNP\Yosemite.aprx")
mp = aprx.listMaps('Yosemite National Park')[0]
for lyr in mp.listLayers():
    if lyr.name == 'Roads':
        if lyr.maxThreshold == 10000:
            lyr.visible = True
        if lyr.maxThreshold == 100000:
            lyr.visible = False
aprx.save()
del aprx
```

arcpy.mp functions

ArcGISProject(aprx_path)

ColorRamp(color_ramp_name, {index})

ConvertLayoutFileToLayout(layout_file)

ConvertWebMapToArcGISProject(webmap_json, {template_page}, {mapframe_name}, {notes_gdb})

CreateWebLayerSDDraft(map_or_layers, out_sddraft, service_name, {server_type}, {service_type}, {folder_name}, {overwrite_existing_service}, {copy_data_to_server}, {enable_editing}, {allow_exporting}, {enable_sync}, {summary}, {tags}, {description}, {credits}, {use_limitations})

LayerFile(layer_file_path)

PDFDocumentCreate(pdf_path)

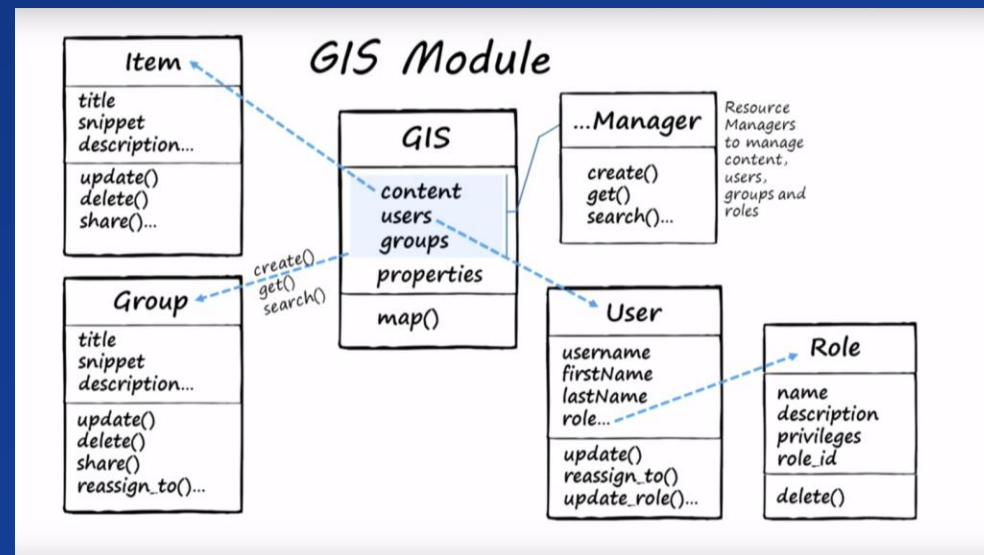
PDFDocumentOpen(pdf_path, {user_password}, {master_password})

Table(table_data_source)

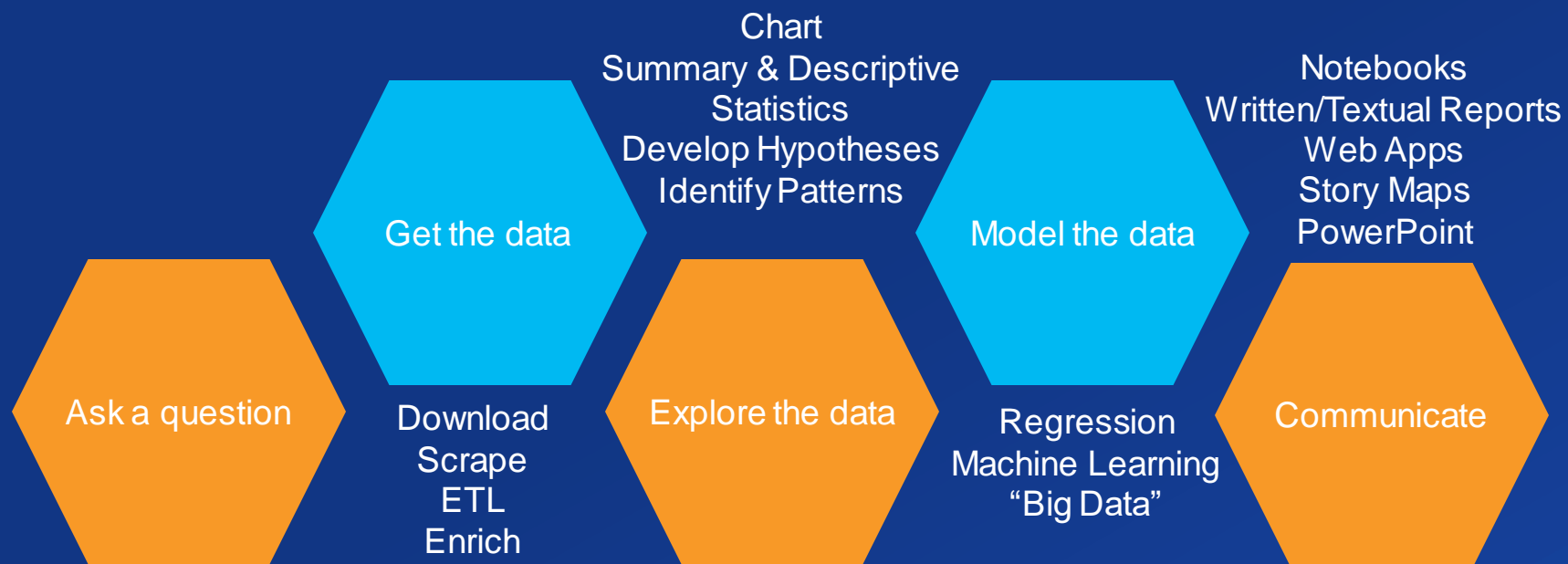
Python

ArcGIS API for Python

- On your **Web GIS** perform:
 - GIS visualization
 - Analysis
 - Spatial data management
- Leverage SciPy ecosystem for
 - Automating workflows
 - Performing repetitive tasks
- Integrate with notebooks for **interactive scripting and visualization**



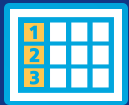
Data Science Process



Get the Data

Non-GIS Data

- Large number of libraries to download, transform, condition, and prepare data
- Generic web libraries (Requests, urllib(2))
- API Specific libraries (Tweepy)
- Ability to scrape and parse existing web sites (Scrapy and BeautifulSoup)



BeautifulSoup



GIS Data

- ArcPy allows native access to all Esri Data formats and the full functionality of ArcGIS Desktop
- ArcGIS API for Python allows for access and interaction with content within your WebGIS



Explore the Data

- **ArcGIS Notebooks** are your friend
- Ability to create **Spatially Enabled DataFrame (SEDF)**
- Core Pandas DataFrame, allows for rapid exploration of data
- Can create maps, charts and a variety of other objects quickly using a common syntax
- Availability
 - ArcGIS Pro Notebooks 2.5+
 - ArcGIS Notebooks for Enterprise 10.7+
 - We plan to open public beta for ArcGIS Online Notebooks in March with a full release later in the year if all goes well

```
ArcGIS | Deep Learning: Downscale climate models (unsaved changes)
File Edit View Insert Cell Kernel Help
+ 90% Run Code

station_GCH_sdf[station_predictand])

Out[31]: RidgeCV(alphas=array([1.000000e-10, 1.66810e-08, 2.78256e-06, 4.64159e-04, 7.74264e-02, 1.29155e+01, 2.15443e+03, 3.59381e+05,
5.99484e+07, 1.00000e+10]),
cv=None, fit_intercept=True, gcv_mode=None, normalize=True,
scoring=None, store_cv_values=True)

In [32]: ax = plt.gca()
ax.plot(ridge.alphas.ravel(),
np.mean(ridge.cv_values_,
axis = 0).ravel())
ax.set_xscale('log')
ax.set_xlim(ax.get_xlim()[::-1]) # reverse axis
plt.xlabel('regularization Strength')
plt.ylabel('CV Error')
plt.title('CV Error vs Alpha')
plt.axis('tight')
plt.show()

CV Error vs Alpha

30
25
20
15
10
5
10^10 10^8 10^6 10^4 10^2 10^0 10^-2 10^-4 10^-6 10^-8 10^-10
Regularization Strength

In [33]: print('Optimal Regularization Parameter = {}'.format(ridge.alpha_))
Optimal Regularization Parameter = 0.000464156883612782

3.1.2 Support Vector Machine for Climate Downscaling

In [34]: svr_rbf = SVR(kernel='rbf', C=1.0, gamma='auto').fit(
station_GCH_sdf[station_predictors],
station_GCH_sdf[station_predictand].values.ravel())

In [35]: svr_rbf.predict(output_GCH_sdf[station_predictors])

Out[35]: array([10.178156, 10.247221, 10.389448, 10.494384, ..., 9.024728, 9.024996, 9.054723, 9.058166])

In [36]: from IPython.display import IFrame
IFrame(src='https://esth.maps.arcgis.com/apps/StorytellingSwipe/'\
'index.html?appid=c42c84c661140f6883f9843c06d1d56e',
width='100%', height='800px')

Out[36]:
```

Explore the Data

Spatially Enabled DataFrame

- **Spatial engine built directly into the DataFrame**
 - Use Pandas operations on attribute and spatial columns
- Read in shapefiles, Pandas DataFrames, feature classes, GeoJSON, Feature Layers
- New SEDF object requires ArcGIS API for Python version 1.5
 - Updated from old SDF object
 - Better memory management
 - Handle larger datasets
 - Pandas advocates as the path forward

```
In [6]: from arcgis import GIS
gis = GIS()
item = gis.content.get("85d0ca4ea1ca4b9abf0c51b9bd34de2e")
flayer = item.layers[0]

# create a Spatially Enabled DataFrame object
sdf = pd.DataFrame.spatial.from_layer(flayer)
sdf.head()
```

Out[6]:

DEFIPS	POP2010	POPULATION	POP_CLASS	RENTER_OCC	SHAPE	ST	STFIPS	VACANT	WHITE
220	39540	40346	6	6563	["x": -12751215.004681978, "y": 4180278.406256...	AZ	04	6703	32367
895	14364	14847	6	1397	["x": -12755627.731115643, "y": 4164465.572856...	AZ	04	1389	12730
030	26265	26977	6	1963	["x": -12734674.294574209, "y": 3850472.723091...	AZ	04	9636	22995
370	52527	55041	7	6765	["x": -12725332.21151233, "y": 4096532.0908223...	AZ	04	9159	47335
470	25505	29767	6	1681	["x": -12770984.257542243, "y": 3826624.133935...	AZ	04	572	16120

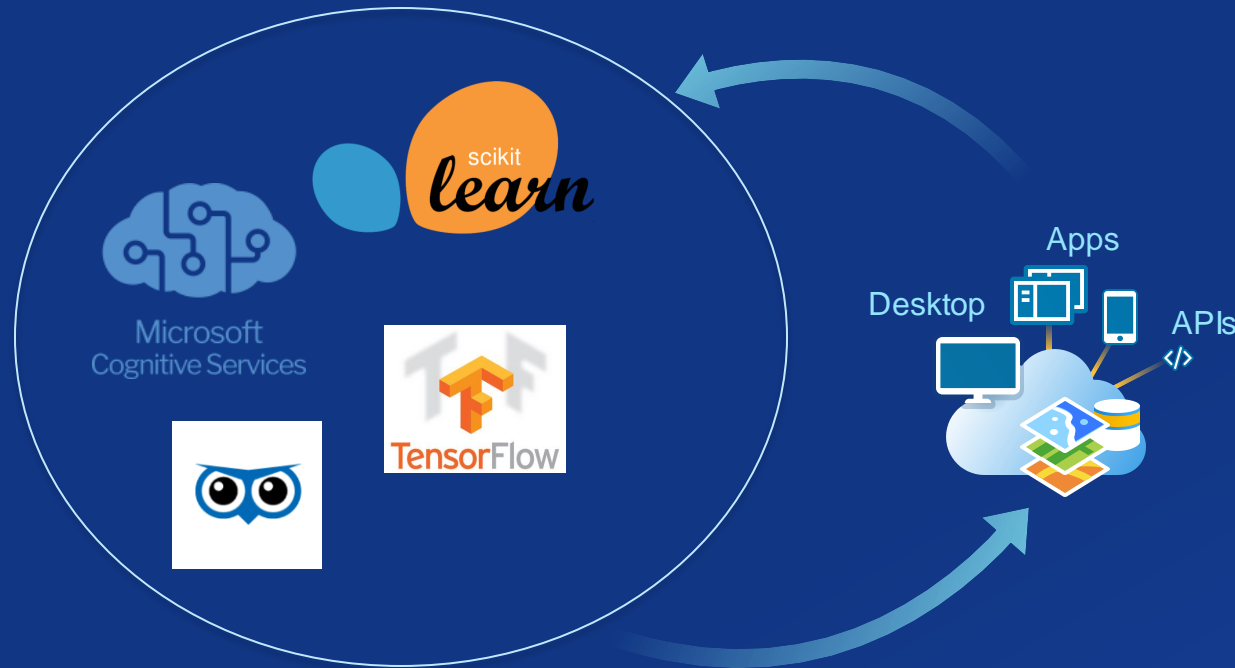
Explore the Data

arcgis.learn Module

- Supports AI/ML workflows within the ArcGIS platform
 - Exporting training data
 - Data prep
 - Model training
 - Model deployment
 - Model inference
- Leverage tools across:
 - ArcGIS API for Python
 - ArcGIS Pro
 - ArcGIS Image Server
 - ArcGIS Notebooks



Modeling the Data



- Many native tools to enable modeling of data (Space-Time Pattern Mining, Density Based Clustering, etc...)
- Integration of popular third-party Machine Learning/Deep Learning libraries
 - Scikit-Learn
 - Tensorflow
 - PyTorch
 - NLTK



Python Demo

ArcGIS Notebooks

Data Engineering

Visualization & Exploration

Spatial Analysis



Resources

Documentation

- [What is ArcPy?](#)
- [Overview of the ArcGIS API for Python](#)
- [Introducing ArcGIS Notebooks](#)
- [Spatial Analysis and Data Science](#)

Community

- [SciPy](#)
- [Python – Numeric and Scientific Libraries](#)

Spatial Statistics GitHub

- [Spatial Statistics Resources](#)

GeoNet

- [Spatial Statistics](#)
- [Python](#)
- [ArcGIS API for Python](#)

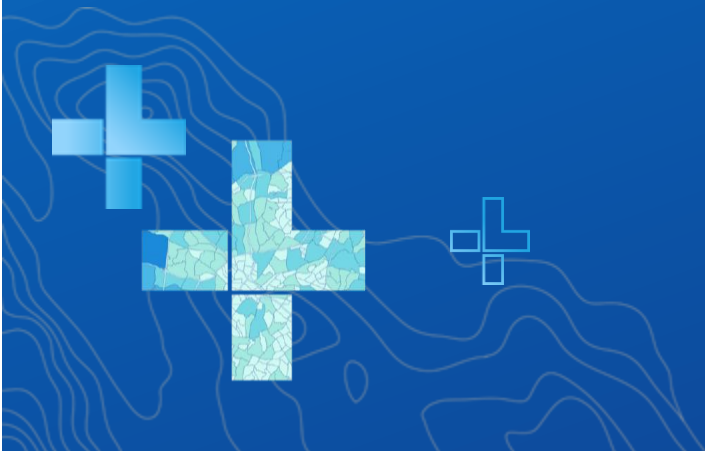


Key Takeaways

- **Knowledge** – Use the best tool for the job - ArcPy and ArcGIS API for Python can help accomplish complex, data science workflows
- **Integration** – ArcGIS is an open platform that supports end-to-end analytic workflows. Leverage third party libraries.
- **Communication** – Harness your portal to deliver information products to your organization/community

Utilizing the R- ArcGIS Bridge

An Introduction



What is R?



- A widely used statistical programming language
 - Linear and non-linear modeling / Classification / Clustering ...
 - Data Cleaning & Transformation
 - Data Visualization
- More than 12,000 Packages
 - General to Specific Use Cases
- Open-Source
 - Universities, Government, Industry
 - Finance, Ecology, Statistics
 - Machine learning, predictive analytics

What is the R-ArcGIS Bridge

- Provides the ability for users to integrate R and ArcGIS functionality
- Open source project, free to download and use

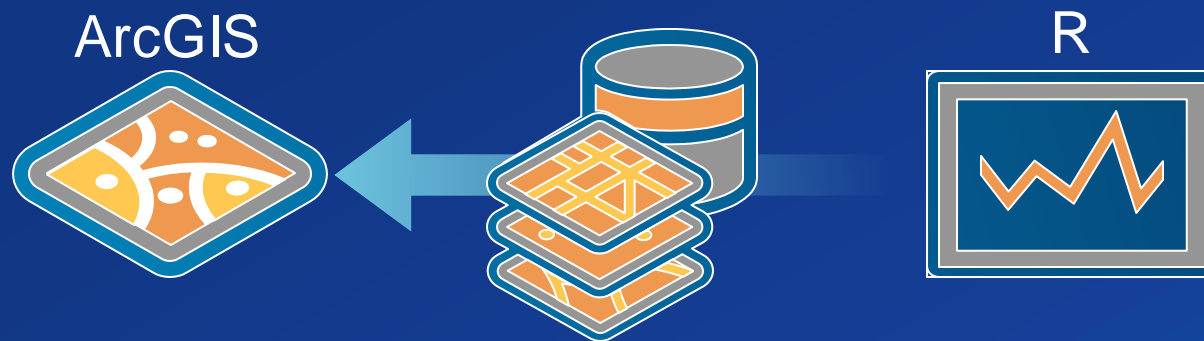


Who Can Use the R-ArcGIS Bridge?

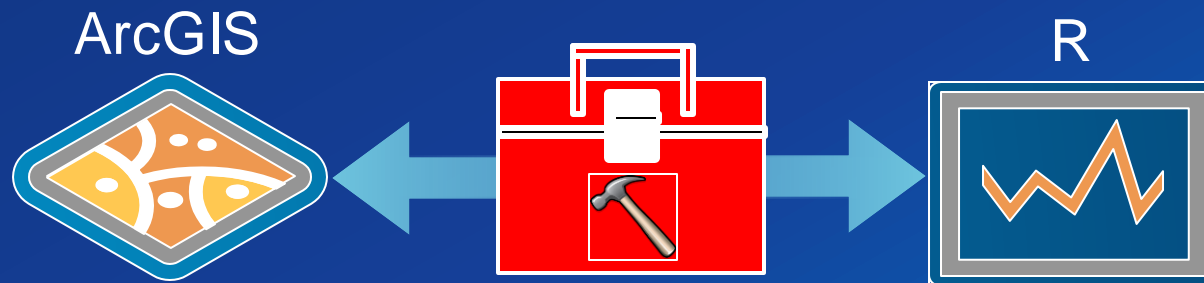
GIS Analyst



Data Scientist



Developers



Version Requirements for the R-ArcGIS Bridge

ArcGIS Pro



1.1 (or later)

ArcMap



10.3.1 (or later)

R

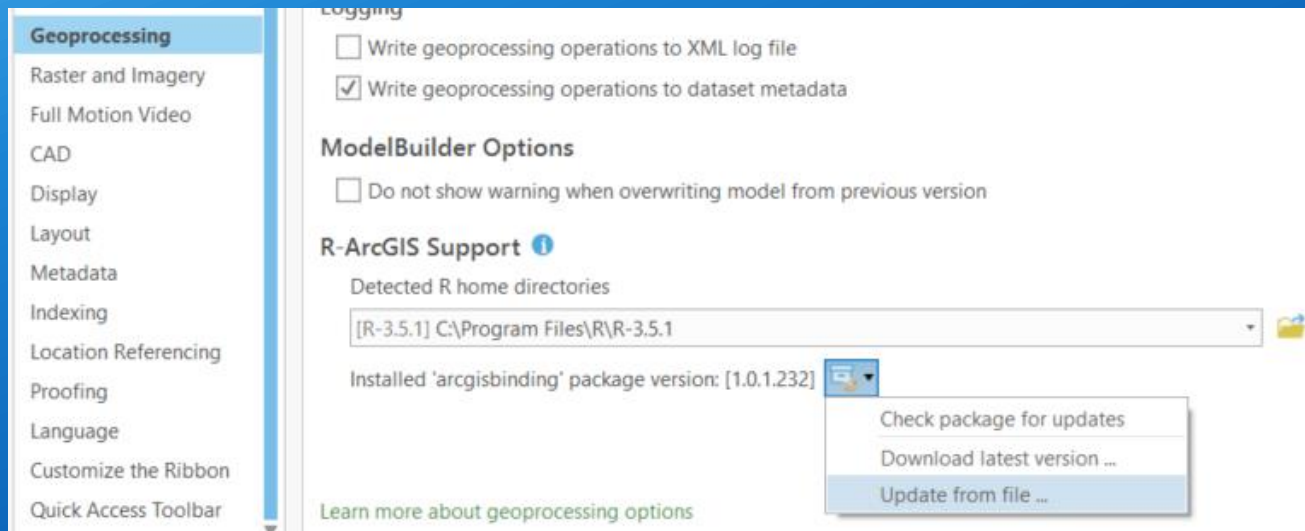


3.2.2 (or later)

RStudio / Notebook



Recommended !



Installing the Bridge / Getting Started ArcGIS Pro *Project Tab*

Vector Support

- Ability to read, write, and modify vector data
- Support / Conversion for key R objects and spatial packages
 - R data frame
 - Compatibility with **sp** package
 - Compatibility with **sf** package
- Customize data manipulations
 - SQL queries / Subsetting
 - Data Clean Up and Engineering
- Maintain spatial geometries when working with **dplyr**

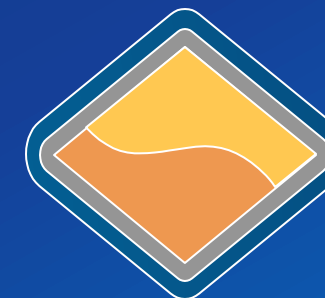
Points



Lines

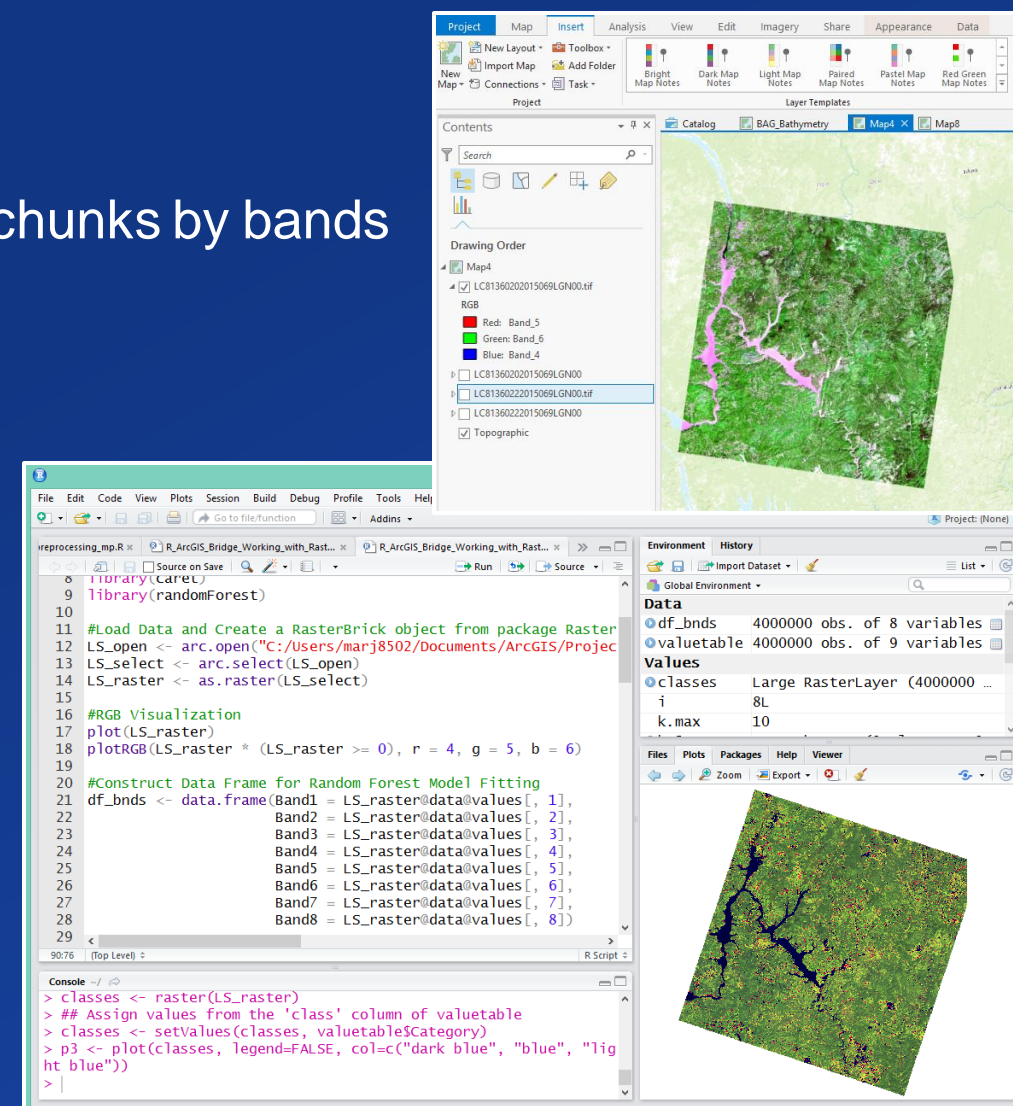


Polygons



Raster Support

- Ability to read and write raster data
 - Handle big data raster data with the ability to read in chunks by bands
 - Compatibility with CRF format and Mosaic Datasets
- Customize selections and subsets
 - Create subsets by bands or pixel rows and columns
 - Resample options available
 - Select desired pixel format for specific analyses

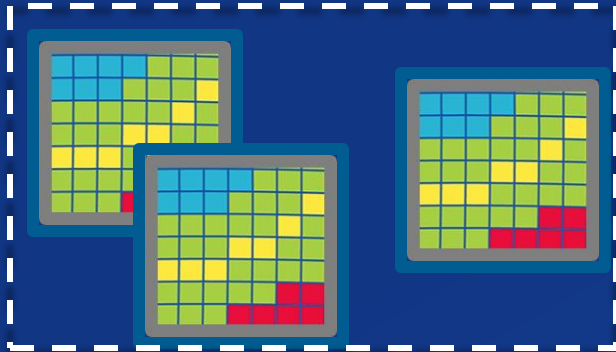


Support for Mosaic Datasets

- Raster data can become a big data problem, quickly

Mosaics: Data structure to store/process rasters in space and time

Mosaics



Time-Series Rasters/Mosaic



Working with Big Data

- Microsoft Open R is a publicly available R-version for big data
- Contains almost all CRAN libraries
- Window-based operations and image operators speed up drastically
- Set-up and Usage from Pro is exactly the same as traditional R





R-ArcGIS Bridge Demo

Predict Home Vacancy Rates in
Washington DC based on Vacant
Home Locations in Baltimore (XY
Coordinates)

Key Takeaways

- ArcGIS is an open platform that supports end to end analytic workflows
- Designed to integrate with your favorite modules in Python and R
- Communicate your results with the community through ArcGIS Online / Enterprise



Resources

Learn More on Using the R-ArcGIS Bridge

Resources from Spatial Statistics Team:

- <https://github.com/R-ArcGIS>

Getting Started:

- [Analyzing Crime Using Statistics and the R-ArcGIS Bridge Learn Lesson](#)
- [Using the R-ArcGIS Bridge Introductory Web Course](#)

Creating R Script Tools:

- [Integrating R Scripts into Geoprocessing Tools Web Course](#)
- [arctgisbinding Package Vignette](#)

Powerful, In-depth Workflows in ArcGIS and R

- [Identify an Ecological Niche for African Buffalo](#)



Relevant Upcoming Sessions: Tuesday February 11

- **Data Science Using ArcGIS Notebooks**

- 150 A | 3:00 – 4:00 pm

- **Machine Learning in ArcGIS**

- 146 C | 3:00 – 4:00 pm

- **Using Deep Learning with Imagery in ArcGIS**

- 144 A | 3:00 – 4:00 pm

- **From Means and Medians to Machine Learning: Spatial Statistics Basics and Innovations**

- 146 C | 4:15 – 5:15 pm

- **Python: Beyond the Basics**

- 146 B | 4:15 – 5:15 pm

Relevant Upcoming Sessions: Wednesday February 12

- **Machine Learning in ArcGIS**
 - 146 C | 8:30 – 9:30 am
- **Using Deep Learning with Imagery in ArcGIS**
 - 144 A | 8:30 – 9:30 am
- **Data Visualization for Spatial Analysis**
 - 146 C | 11:00 am – 12:00 pm
- **From Means and Medians to Machine Learning: Spatial Statistics Basics and Innovations**
 - 146 C | 1:30 – 2:30 pm
- **Geospatial Deep Learning with arcgis.learn**
 - 101 | 1:30 – 2:30 pm

Relevant Upcoming Sessions: Wednesday February 12

- **Spatial Data Mining: Cluster Analysis and Space-Time Analysis**
 - 146 C | 2:45 – 3:45 pm
- **Python: Building Geoprocessing Tools**
 - 147 B | 4:00 – 5:00 pm
- **ArcGIS API for Python: Getting to Know Pandas and the Spatial Enabled DataFrame**
 - 146 C | 5:15 – 6:15 pm
- **Python: Map Automation in ArcGIS Pro**
 - 102 A | 5:15 – 6:15 pm
- **The Forest for the Trees: Making Prediction Using Forest-Based Classification and Regression**
 - 146 C | 5:15 – 6:15 pm

Print Your Certificate of Attendance

Print Stations Located in 150 Concourse Lobby

Tuesday

12:30 pm – 6:30 pm
Expo
Hall B

5:15 pm – 6:30 pm
Expo Social
Hall B

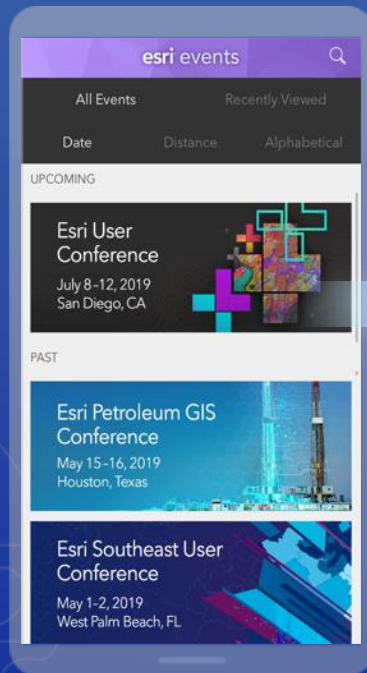
Wednesday

10:45 am – 5:15 pm
Expo
Hall B

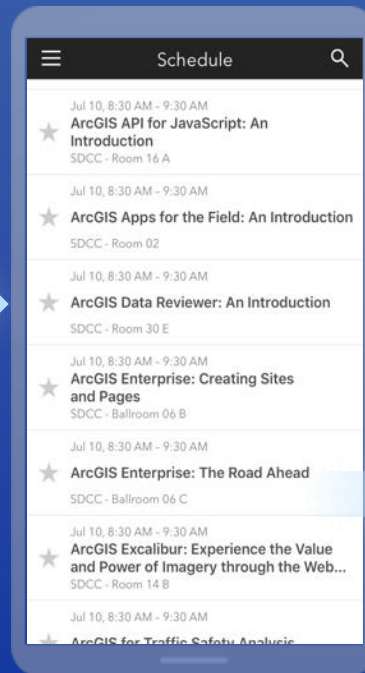
6:30 pm – 9:30 pm
Networking Reception
Smithsonian National Museum
of Natural History

Please Share Your Feedback in the App

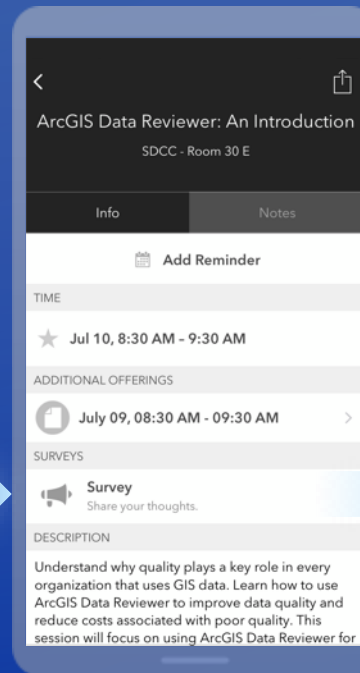
Download the Esri Events app and find your event



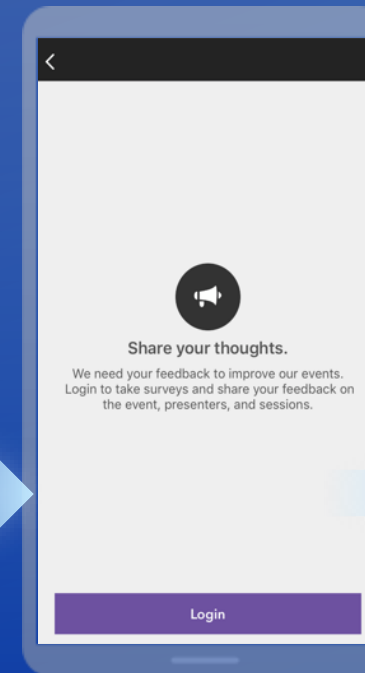
Select the session you attended



Scroll down to "Survey"



Log in to access the survey



Complete the survey and select "Submit"

