

# Building “Geographic Data Science...”

Dani Arribas-Bel [@darribas]



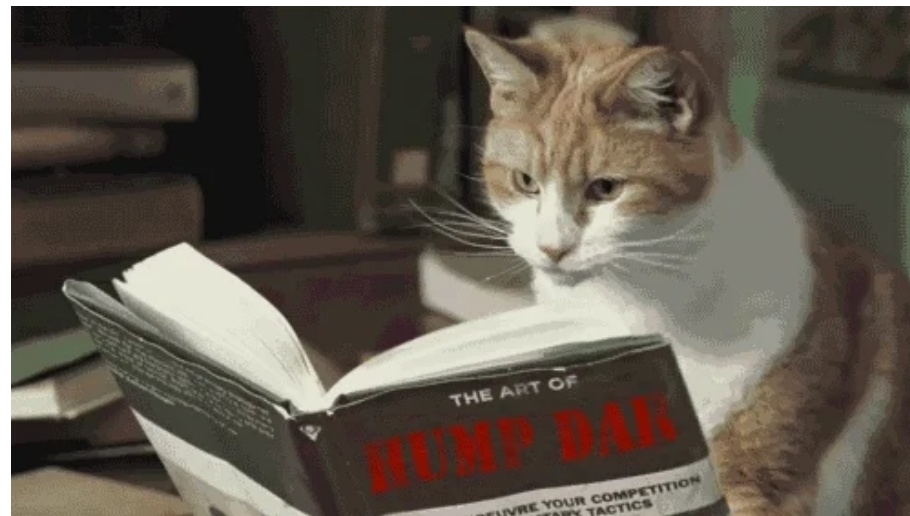
UNIVERSITY OF  
LIVERPOOL

**The  
Alan Turing  
Institute**



Geographic  
Data Science  
Lab

# We have a book!



via GIPHY

# Almost...




via GIPHY

Coming in 2021 but...

# Coming in 2021 but...

... you can already:

- <https://geographicdata.science>
-  launch binder
- <https://github.com/gdsbook/book>



# The Authors



@sreyog

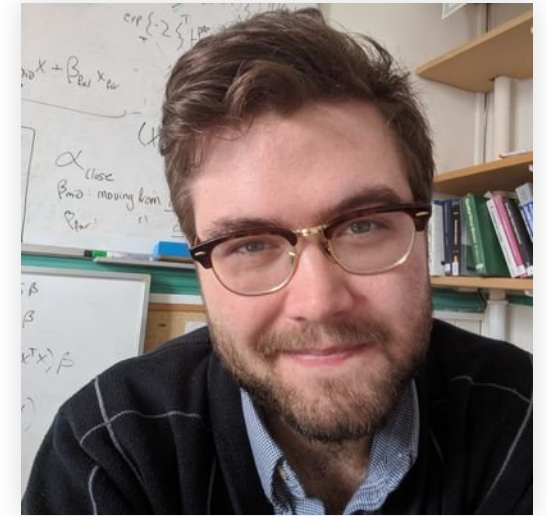
Serge Rey



@darribas

Dani

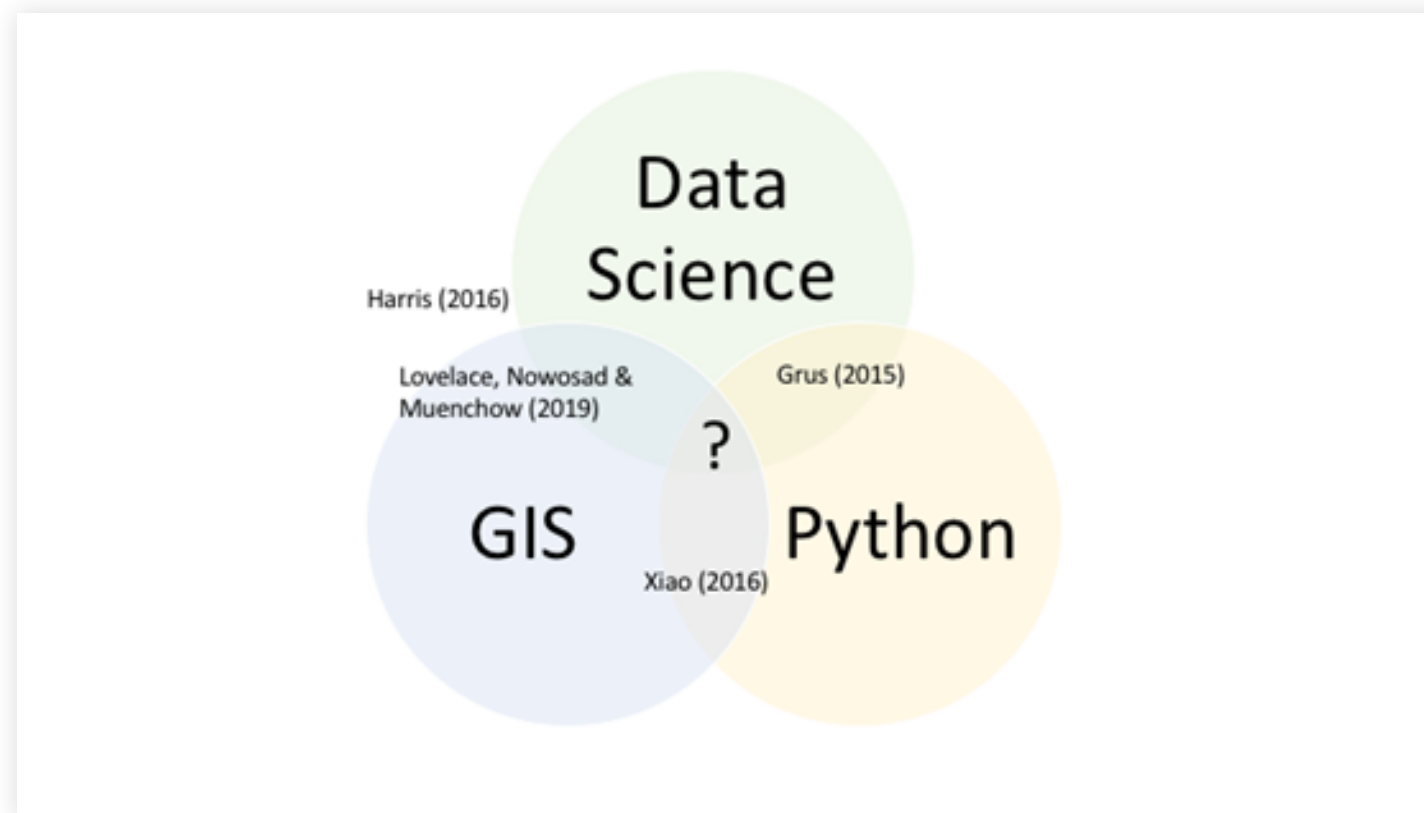
Arribas-Bel



@levijohnwolf

Levi Wolf

# The Book



# This Talk

- Why
- What
- How

Why



Data, data, data

Data Science


...

**It's called  
GEOGRAPHIC Data  
Science!!!**



# Geographic Data Science

What



Geographic Data Science with Python

Search this book...

- Home
- PREFACE
- Table of Contents**
- References
- PART I - BUILDING BLOCKS**
  - Overview
  - Geospatial Computational Environment
  - Geographic thinking for data scientists
  - Spatial Data Processing
  - Spatial Weights
- PART II - SPATIAL DATA ANALYSIS**
  - Overview
  - Choropleth Mapping
  - Global Spatial Autocorrelation
  - Local Spatial Autocorrelation
  - Point Pattern Analysis
- PART III - ADVANCED TOPICS**
  - Overview

# Table of Contents

## Part I: Building Blocks

- [Geospatial Computational Environment](#)
- [Spatial data](#)
- [Spatial data processing](#)
- [Spatial weights](#)

## Part II: Spatial Data Analysis

- [Choropleth Mapping](#)
- [Spatial Autocorrelation](#)
- [Local Spatial Autocorrelation](#)
- [Point Pattern Analysis](#)

## Part III: Advanced Topics

- [Spatial Inequality](#)
- [Clustering and Regionalization](#)
- [Spatial Regression](#)
- [Spatial Feature Engineering](#)

[<< Home](#) [References >>](#)

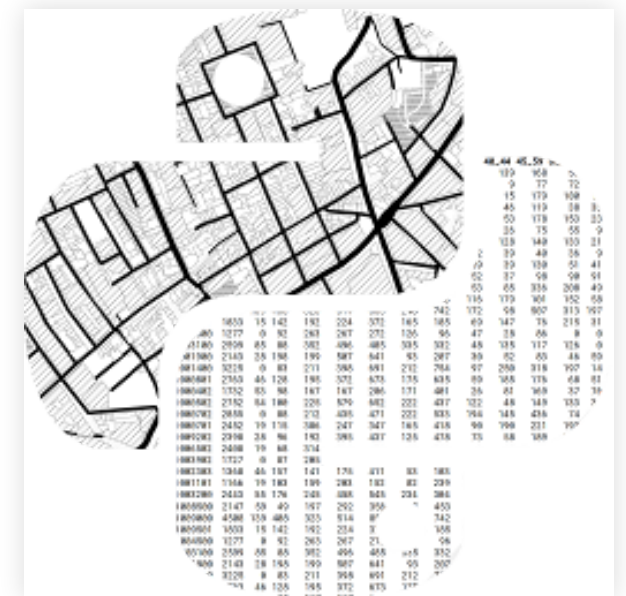
By Sergio J. Rey, Dani Arribas-Bel, Levi J. Wolf  
© Copyright 2020.



This work is licensed under a [Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License](#).

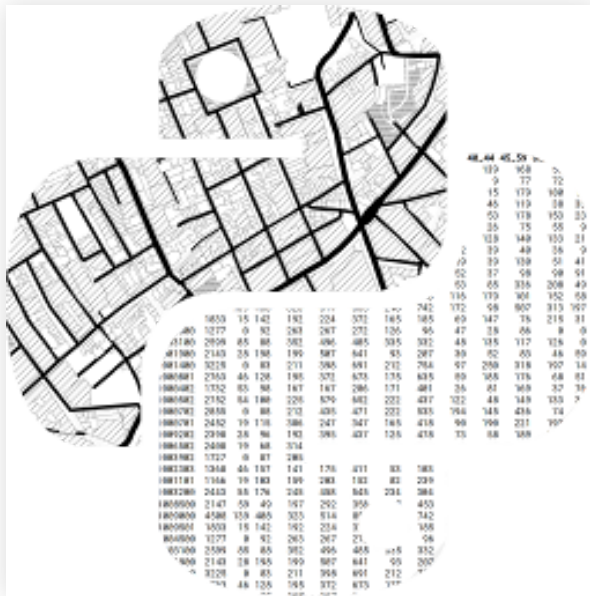
# Building Blocks

- Computational Environment
- Geographical Thinking
- Spatial Data
- Spatial Weights



# Fundamentals

- Choropleths
- Spatial Autocorrelation
- Local Spatial Autocorrelation
- Point Patterns







# Bonus: Datasets

### AirBnb

```
import pandas as pd
import requests
import numpy as np
import json
from tqdm import tqdm
import time
```

Download files

- Download dataset files

```
url = "https://data.airbnb.com/airbnb-properties/v1/states/us"
headers = {"User-Agent": "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7; rv:68.0) Gecko/20100801 Firefox/68.0"}
response = requests.get(url, headers=headers)
df = pd.DataFrame(response.json())
df.to_csv('airbnb_us.csv', index=False)
```

Download dataset files with proxy

```
url = "https://data.airbnb.com/airbnb-properties/v1/states/us"
headers = {"User-Agent": "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7; rv:68.0) Gecko/20100801 Firefox/68.0"}
response = requests.get(url, headers=headers, proxies={"http": "http://127.0.0.1:8080", "https": "https://127.0.0.1:8080"})
df = pd.DataFrame(response.json())
df.to_csv('airbnb_us.csv', index=False)
```

### Airports

```
import pandas as pd
import requests
import numpy as np
import json
from tqdm import tqdm
import time
```

```
url = "https://data.airbnb.com/airbnb-properties/v1/states/us"
headers = {"User-Agent": "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7; rv:68.0) Gecko/20100801 Firefox/68.0"}
response = requests.get(url, headers=headers)
df = pd.DataFrame(response.json())
df.to_csv('airbnb_us.csv', index=False)
```

scale	radius	type	name	abbr	location	gpc_code	lat	lon
0	9	Airport	small	Sherwood	LLH	terminal	VLD	LLH

### Brexit

Brexit dataset

This dataset contains results for the Brexit vote at the local authority district and administrative boundaries.

[brexit\\_vote.csv](#)

- Source: Electoral Commission
- File

[http://www.electoralcommission.org.uk/\\_data/assets/file/0114/21113/BV-referendum-result-data.csv](http://www.electoralcommission.org.uk/_data/assets/file/0114/21113/BV-referendum-result-data.csv)

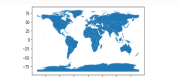
Processing/proxy processing was required for this dataset, see original source for additional information.

Local Authority District boundaries

### Countries

```
import pandas as pd
import requests
import numpy as np
import json
from tqdm import tqdm
import time
```

```
url = "https://data.airbnb.com/airbnb-properties/v1/states/us"
headers = {"User-Agent": "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7; rv:68.0) Gecko/20100801 Firefox/68.0"}
response = requests.get(url, headers=headers)
df = pd.DataFrame(response.json())
df.to_csv('airbnb_us.csv', index=False)
```



### H3 Grid

Build a H3 grid for the San Diego region

Infrastructure

To create a container that includes all the following on a file called `h3grid.py`:

```
import h3
import requests
import pandas as pd
import numpy as np
import json
import time
```

And include the container by running the following from the same folder where the file is stored:

```
docker build -t h3grid .
```

### Mexico

This dataset contains Decadal GDP (Exports for Mexico states, from 1940-2000).

- Source: FRED, data source: FRED, data source: FRED
- File

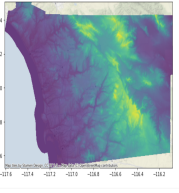
<https://fred.stlouisfed.org/series/USGDP>

Processing/proxy processing was required for this dataset, see original source for additional information.

By Sergio J. Rey, Don Anbar-Bell, Lev I. Wolf  
© Copyright 2020.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

```
r = rasterio.open("usstate_us.tif")
r.crs = "EPSG:4326"
r.bounds = (min(r.bounds[0], r.bounds[2]), max(r.bounds[0], r.bounds[2]), min(r.bounds[1], r.bounds[3]), max(r.bounds[1], r.bounds[3]))
```



### Texas

This dataset includes geometries for Texas counties.

- Source: Geographic Data Science with Python, SciPy 16
- File

<https://github.com/robertocarreras/geojson>

Processing/proxy processing was required for this dataset, see original source for additional information.

By Sergio J. Rey, Don Anbar-Bell, Lev I. Wolf  
© Copyright 2020.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

### Tokyo Photographs

```
url = "https://data.airbnb.com/airbnb-properties/v1/states/us"
headers = {"User-Agent": "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7; rv:68.0) Gecko/20100801 Firefox/68.0"}
response = requests.get(url, headers=headers)
df = pd.DataFrame(response.json())
df.to_csv('airbnb_us.csv', index=False)
```

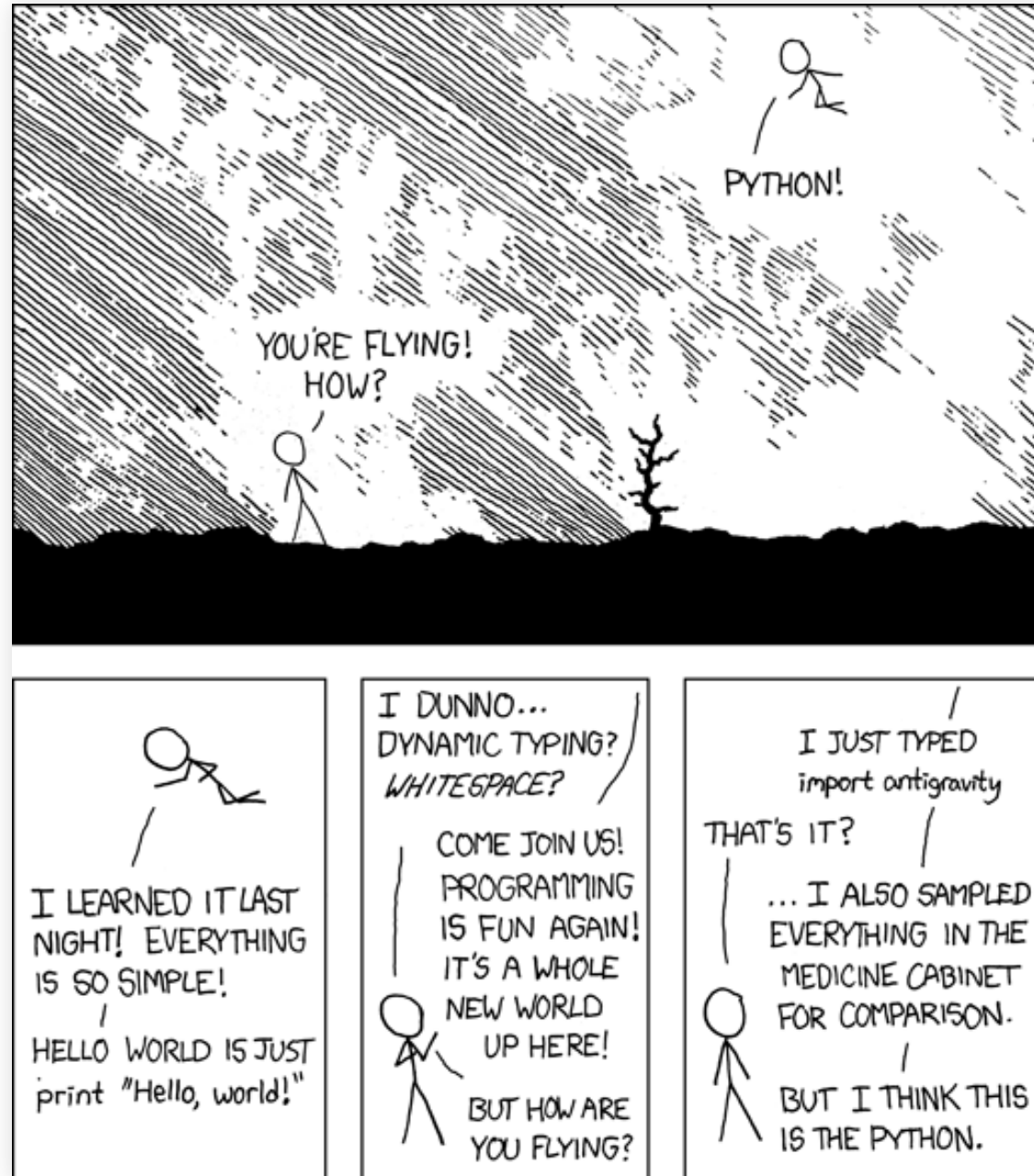
Randomly subsampling

```
r = rasterio.open("usstate_us.tif")
r.crs = "EPSG:4326"
r.bounds = (min(r.bounds[0], r.bounds[2]), max(r.bounds[0], r.bounds[2]), min(r.bounds[1], r.bounds[3]), max(r.bounds[1], r.bounds[3]))
```

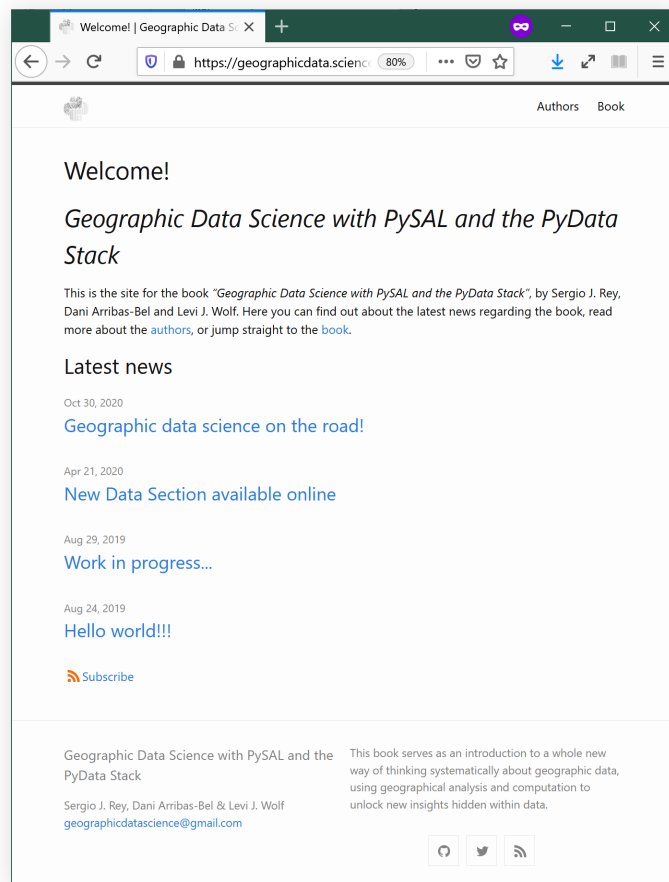
Reproject XY coordinates in separate

**How**

# Python



# Radically Open



>Welcome! | Geographic Data Science

<https://geographicdata.science>

Authors Book

## Welcome!

### Geographic Data Science with PySAL and the PyData Stack

This is the site for the book "Geographic Data Science with PySAL and the PyData Stack", by Sergio J. Rey, Dani Arribas-Bel and Levi J. Wolf. Here you can find out about the latest news regarding the book, read more about the [authors](#), or jump straight to the [book](#).

#### Latest news

Oct 30, 2020  
[Geographic data science on the road!](#)

Apr 21, 2020  
[New Data Section available online](#)

Aug 29, 2019  
[Work in progress...](#)

Aug 24, 2019  
[Hello world!!!](#)

[Subscribe](#)

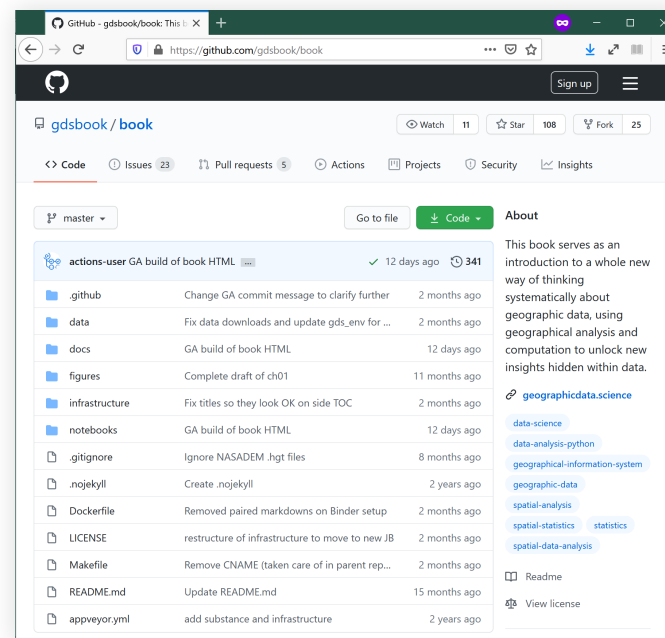
---

Geographic Data Science with PySAL and the PyData Stack

This book serves as an introduction to a whole new way of thinking systematically about geographic data, using geographical analysis and computation to unlock new insights hidden within data.

Sergio J. Rey, Dani Arribas-Bel & Levi J. Wolf  
[geographicdatascience@gmail.com](mailto:geographicdatascience@gmail.com)

[🔄](#) [🐦](#) [📡](#)



gdsbook / book

Code Issues (23) Pull requests (5) Actions Projects Security Insights

master

actions-user GA build of book HTML 12 days ago 341

- .github Change GA commit message to clarify further 2 months ago
- data Fix data downloads and update gds\_env for ... 2 months ago
- docs GA build of book HTML 12 days ago
- figures Complete draft of ch01 11 months ago
- infrastructure Fix titles so they look OK on side TOC 2 months ago
- notebooks GA build of book HTML 12 days ago
- .gitignore Ignore NASADEM .hgt files 8 months ago
- .nojekyll Create .nojekyll 2 years ago
- Dockerfile Removed paired markdowns on Binder setup 2 months ago
- LICENSE restructure of infrastructure to move to new JB 2 months ago
- Makefile Remove CNAME (taken care of in parent rep... 2 months ago
- README.md Update README.md 15 months ago
- appveyor.yml add substance and infrastructure 2 years ago

About

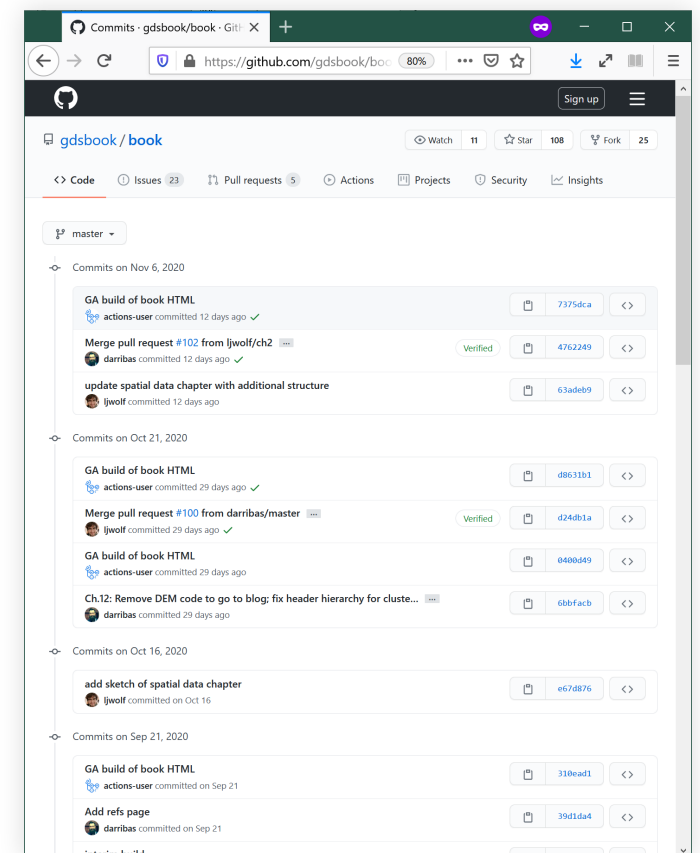
This book serves as an introduction to a whole new way of thinking systematically about geographic data, using geographical analysis and computation to unlock new insights hidden within data.

[geographicdata.science](https://geographicdata.science)

- data-science
- data-analysis-python
- geographical-information-system
- geographic-data
- spatial-analysis
- spatial-statistics
- statistics
- spatial-data-analysis

Readme

View license



Commits - gdsbook/book

Code Issues (23) Pull requests (5) Actions Projects Security Insights

master

Commits on Nov 6, 2020

- GA build of book HTML actions-user committed 12 days ago ✓ 7375dca
- Merge pull request #102 from ljwolf/ch2 darribas committed 12 days ago ✓ Verified 4762249
- update spatial data chapter with additional structure ljwolf committed 12 days ago 63a4e9

Commits on Oct 21, 2020

- GA build of book HTML actions-user committed 29 days ago ✓ d8c31b1
- Merge pull request #100 from darribas/master ljwolf committed 29 days ago ✓ Verified d24d1a
- GA build of book HTML actions-user committed 29 days ago 0480d49
- Ch.12: Remove DEM code to go to blog: fix header hierarchy for cluste... darribas committed 29 days ago 6bbfacb

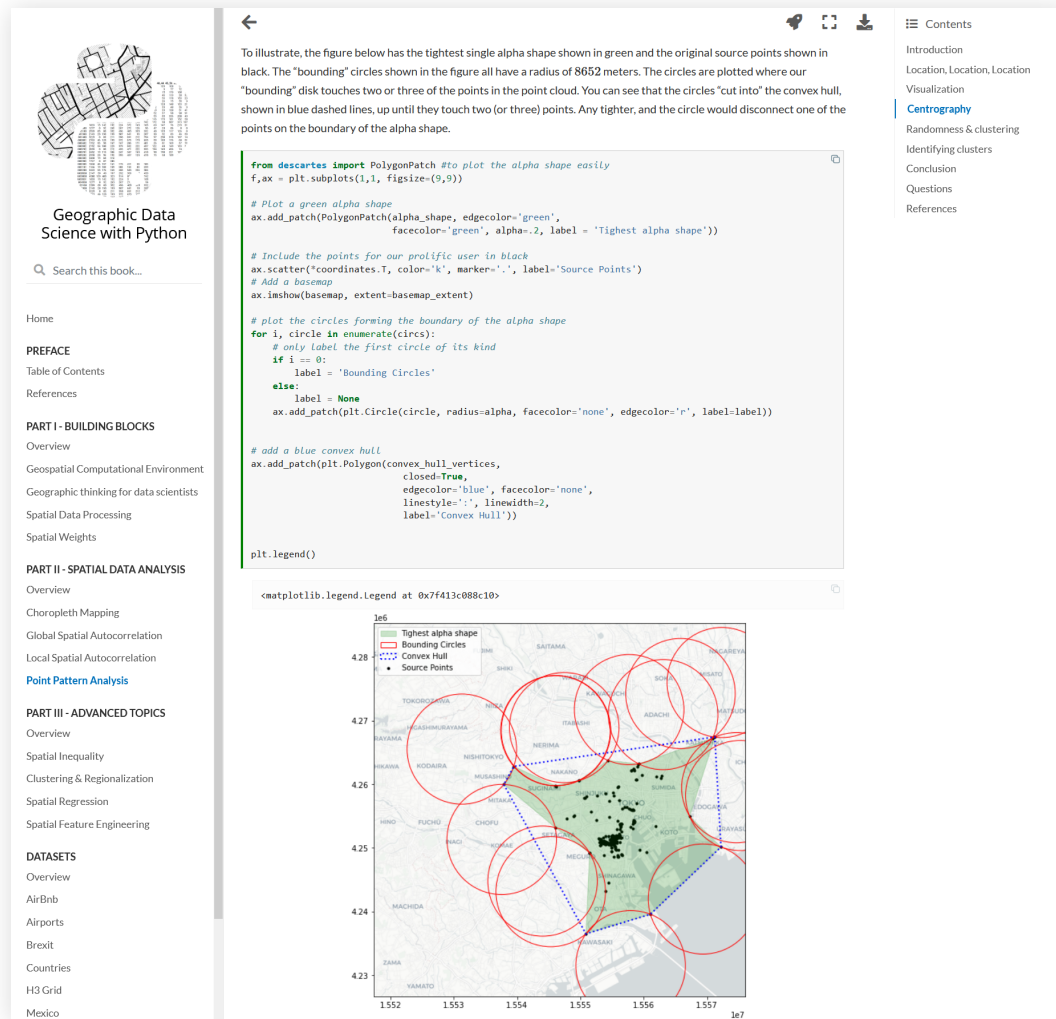
Commits on Oct 16, 2020

- add sketch of spatial data chapter ljwolf committed on Oct 16 e67db76

Commits on Sep 21, 2020

- GA build of book HTML actions-user committed on Sep 21 318ead1
- Add refs page darribas committed on Sep 21 39d1da4
- interim build

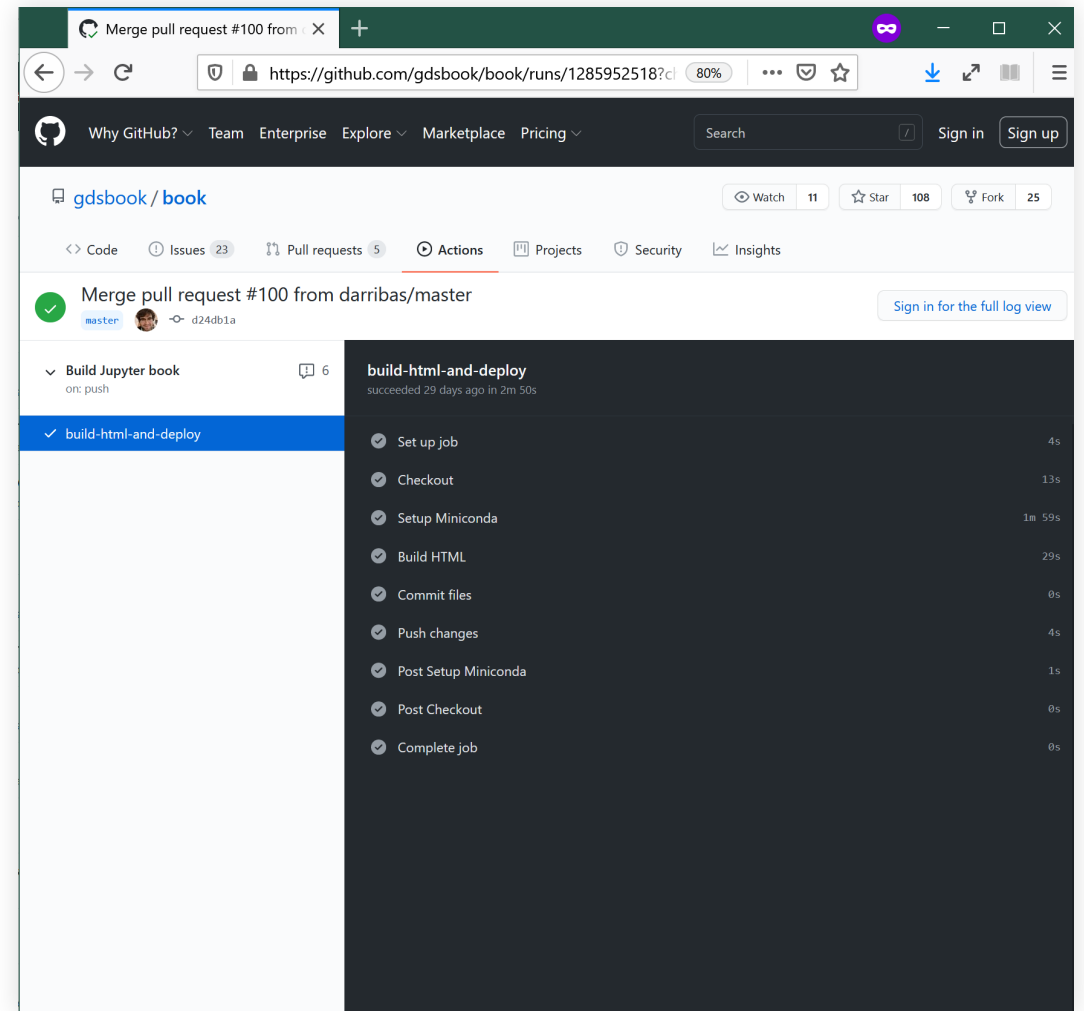
# Code as text; text as code



The screenshot shows a Jupyter Notebook page from the book "Geographic Data Science with Python". The notebook contains Python code for calculating an alpha shape and its convex hull. The code uses the following logic:

- Imports `descartes` for `PolygonPatch`.
- Plots source points as black dots on a map.
- Calculates the "Tightest alpha shape" as a green filled polygon.
- Generates "Bounding Circles" as red circles centered on each source point.
- Calculates the "Convex Hull" as a blue dashed line.

The visualization below the code shows a map of a city area with a green shaded region representing the alpha shape, red circles representing bounding circles, and a blue dashed line representing the convex hull. A legend in the top-left corner of the plot identifies these elements.



The screenshot displays a GitHub Actions workflow run for a pull request. The workflow is named `build-html-and-deploy` and is triggered on a push to the `master` branch. The workflow consists of the following steps:

- `Set up job` (4s)
- `Checkout` (13s)
- `Setup Miniconda` (1m 59s)
- `Build HTML` (29s)
- `Commit files` (0s)
- `Push changes` (4s)
- `Post Setup Miniconda` (1s)
- `Post Checkout` (0s)
- `Complete job` (0s)

The workflow is shown as successful, with a green checkmark and the text "succeeded 29 days ago in 2m 50s".

# Try it out!!!

The screenshot shows a web browser window with the following elements:

- Browser Tab:** "Spatial Feature Engineering — X"
- Address Bar:** "https://geographicdata.science" with a 90% zoom level.
- Page Header:** "Spatial Feature Engineering" with navigation icons (back, forward, refresh, search, share, print, menu).
- Left Sidebar:**
  - Logo: "Geographic Data Science with Python"
  - Search: "Search this book"
  - Home
  - PREFACE
  - Table of Contents
  - References
- Main Content:**
  - Section Header: "Spatial Feature Engineering"
  - Text: "In machine learning and data science, we are often equipped with *tons* of data. Indeed, given the constellation of packages to query data services, free and open source data sets, and the rapid and persistent collection of geographical data, there is simply too much data to even represent coherently in a single, tidy fashion. However, we often need to be able to construct useful *features* from this rich and deep sea of data."
  - Text: "Where data is available, but not yet directly *usable*, *feature engineering* helps to construct useful data for modelling a..."
- Right Sidebar:**
  - Contents
    - What is spatial feature engineering?
    - Feature Engineering Using Map Matching
    - Feature Engineering using Map Synthesis
    - Conclusion
    - Questions
- Bottom:** A URL bar showing "https://mybinder.org/v2/gh/gdsbook/book/master?urlpath=lab/tree/notebooks/12\_feature\_engineering.ipynb".

[PDF version of these slides]



GDS - The Book is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.