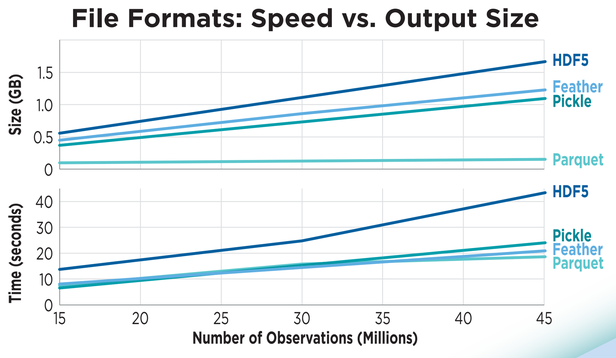
Building the Foundations for AI Use in Regional Statistics

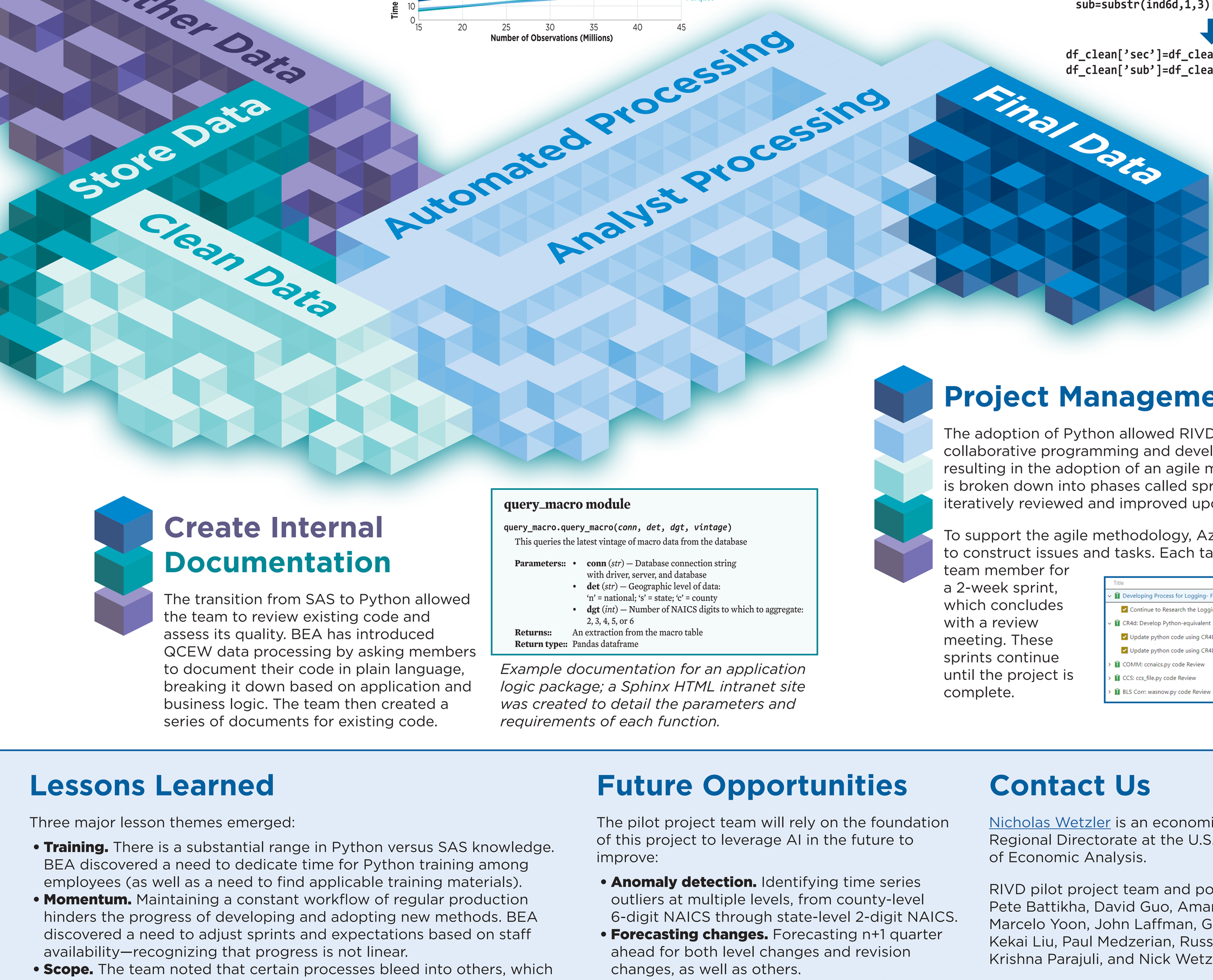
The U.S. Bureau of Economic Analysis (BEA) has adopted Python as the programming language of choice for production of its estimates. BEA's Regional Income and Value Added Division (RIVD) pilot project team has taken this opportunity to modernize the SAS-based process into a Python-based process. Generating estimates requires two main aspects: **automated processing** and **analyst processing**. The goal is to aid BEA analysts in detecting and correcting anomalies and invalid data points.

Bureau of Economic Analysis
U.S. DEPARTMENT OF COMMERCE

Data Storage and Analysis

RIVD's data are currently stored in an extensive system of SAS output files which are used by analysts in their cleaning process, before exporting to central systems. The team tested several storage options—comparing these file-based systems to an SQL server database—and explored several analysis packages. Major decision factors were speed, concurrency handling, and integration. After extensive testing, the team decided to stand up a SQL server database and use parquet for read-only storage.





Code Development

The team adopted a functional programming approach for code development, which included building a library of functions in Python. This library will serve as a foundation as more elements are added to the production system. The current functions center around:

- Reading in data from fixed-length string text format
- Parsing data strings into designated variables
- Cleaning and validating the data
  - » Correct outliers and invalids
  - » Drop invalids
- Storing the data in an SQL server database

```
sec=substr(ind6d,1,2);
  sub=substr(ind6d,1,3)||put(ind6d,$naics2d.);
```

```
df_clean['sec']=df_clean['ind6d'].str[0:2]
df_clean['sub']=df_clean['ind6d'].str[0:3]
```

Project Management

The adoption of Python allowed RIVD to experiment with collaborative programming and development techniques, resulting in the adoption of an agile methodology. Work is broken down into phases called sprints, where it is iteratively reviewed and improved upon.

To support the agile methodology, Azure DevOps is used to construct issues and tasks. Each task is assigned to a team member for a 2-week sprint, which concludes with a review meeting. These sprints continue until the project is complete.

Title	Assigned To	State
Developing Process for Logging- First Pass		Done
Continue to Research the Logging process	Fanning, Garrett	Done
CR4d: Develop Python-equivalent		Done
Update python code using CR4D.sas (python programmer)	Wetzler, Nich...	Done
Update python code using CR4D.sas (SAS programmer)	Moncrief, Rus...	Done
COMM: ccnaics.py code Review		Done
CCS: ccs_file.py code Review		Doing
BLS Corr: wasnow.py code Review		Done

Create Internal Documentation

The transition from SAS to Python allowed the team to review existing code and assess its quality. BEA has introduced QCEW data processing by asking members to document their code in plain language, breaking it down based on application and business logic. The team then created a series of documents for existing code.

query_macro module

`query_macro.query_macro(conn, det, dgt, vintage)`

This queries the latest vintage of macro data from the database

Parameters::
- **conn** (*str*) — Database connection string with driver, server, and database
- **det** (*str*) — Geographic level of data: 'n' = national; 's' = state; 'c' = county
- **dgt** (*int*) — Number of NAICS digits to which to aggregate: 2, 3, 4, 5, or 6

Returns:: An extraction from the macro table
Return type:: Pandas dataframe

Example documentation for an application logic package; a Sphinx HTML intranet site was created to detail the parameters and requirements of each function.

Lessons Learned

Three major lesson themes emerged:

- **Training.** There is a substantial range in Python versus SAS knowledge. BEA discovered a need to dedicate time for Python training among employees (as well as a need to find applicable training materials).
- **Momentum.** Maintaining a constant workflow of regular production hinders the progress of developing and adopting new methods. BEA discovered a need to adjust sprints and expectations based on staff availability—recognizing that progress is not linear.
- **Scope.** The team noted that certain processes bleed into others, which revealed how complex the transition process will be. BEA determined that entire interconnected systems need to be transitioned.

Future Opportunities

The pilot project team will rely on the foundation of this project to leverage AI in the future to improve:

- **Anomaly detection.** Identifying time series outliers at multiple levels, from county-level 6-digit NAICS through state-level 2-digit NAICS.
- **Forecasting changes.** Forecasting n+1 quarter ahead for both level changes and revision changes, as well as others.
- **Imputation.** Automating the process for individual analyst imputations.

Contact Us

Nicholas Wetzler is an economist in the Regional Directorate at the U.S. Bureau of Economic Analysis.

RIVD pilot project team and poster contributers: Pete Battikha, David Guo, Amanda Budny, Marcelo Yoon, John Laffman, Garrett Fanning, Kekai Liu, Paul Medzerian, Russell Moncrief, Krishna Parajuli, and Nick Wetzler