Session Overview

• Update on modernizing BEA’s Data Science strategy and implementation

• Data Science Spotlights:
  o Improved linking of firms to outside datasets for Direct Investment
  o Building a Python library to easily use data from BEA’s API
  o Forecasting Regional QCEW

• Comments from Discussants

• Q&A
Opportunity: The increased availability of Big Data, sophisticated analysis techniques, improved tools, and Data Science popularity present a large opportunity to further Mission goals.

Mission goals: Data science touches on each of BEA’s four mission goals—(1) accuracy and reliability, (2) relevance, (3) customer service, (4) operational excellence.
Data Science: Slices of Success

Outlier Detection

Health Statistics

“Advance” Services Estimates

COVID-19 Consumer Spending Snapshots
Data Science: Operational Framework

[Open] Executive-Level Office
• Chief Data Scientist and staff
• Oversees execution of BEA’s data science strategy

Data Science Coordinating Team (DSCT)
• Reps from BEA’s economic and support areas
• Serves as a liaison between new executive-level office and other work areas

Desktop Applications Implementation Team (DAIT)
• Existing groups and teams whose work aligns with data science strategy
• Uses existing frameworks, programs, and infrastructure to execute portions of BEA’s data science strategy

Other Cross-Directorate Groups:
Research Coordination Committee
Software User Groups
Workforce Development Team
• Developing a BEA Data Science Curriculum and Toolkit
  o Develop draft personas and pilot strategy, working with DAIT
  o Sponsor Machine Learning for Economists training
  o Continue developing training strategy with DAIT

• Developing a BEA Data Science Project Model as a proof of concept
  o Explore project management structures, including Git and Agile/DevOps
  o Pilot projects with DAIT
Desktop Application Modernization Summary

Goal
• Update software, methods, and development frameworks to the latest best practices
• Facilitate collaboration around the bureau via common software
• Reduce onboarding time and cost
• Implement technology supportive of data science

Process
• Review current software profile
• Identify common calculations and processes
• Research alternatives
• Evaluate software against 15 criteria

Recommendation
• In every BEA program area, Python should be adopted as the language of choice for all code involved in the production of statistical products
Desktop Application Implementation Team Overview

**Infrastructure**
Build an IT infrastructure to support Python that is flexible and provides a secure environment.

**Training**
Establish a high-quality training program for Python that uses a combination of modalities to develop staff skills.

**Workforce Planning**
Formulate a staffing plan to determine the best mix of resources, contracted staff, and career staff to support Python.

**Pilot Projects**
Identify and launch pilot projects to learn from hands-on experience and identify resource requirements.

**Common Code**
Identify and develop a shared library of applications to support a one-BEA approach to statistical production.
Project Roadmap

FY22
- Project Kickoff
- Analyze and assess workforce
- Develop training curriculum
- Develop initial infrastructure
- Execute software and/or support purchases
- Solicit and launch pilot projects
- Create governance framework

FY23
- Complete Pilot Projects
- Adjust infrastructure as necessary
- Develop inventory of applications to migrate to Python
- Mature the governance and best practices
- Develop priority common code
- Determine which software to be kept on the BEA IT platform.

FY24-FY26
- Coordinate the review of existing code to build recode/refactor/rewrite plan
- Execute the migration of existing production code
Spotlights
Spotlight: Entity Matching with BEA Direct Investment Survey Data and Compustat
Improving Accuracy in International

• International’s USDIA and FDIUS surveys contain data that is currently hard to validate
  - Financial and operating data reported on annual and benchmark surveys
  - Income and earnings data reported on quarterly surveys

• One external data source for validation is Compustat, which provides quarterly and annual data on balance sheets, income statements, and cash flow for public companies in the US.
Linking Compustat Data to BEA Data

• Many DI survey respondents cannot be matched to Compustat entities
  o Often because many respondents are private and/or relatively small

• Matching has focused on entities with *outward* DI because a larger proportion of them are public and relatively large
Matching Procedures

• Matching variables: IRS Employer Identification Number (EIN), Company name, Address, Website address, Phone number

• Standardize matching variables
  o Drop special characters (&-,.,%,#)
  o From company names, drop common words and abbreviations (“Inc.”, “Limited”)
  o In address, standardize common abbreviations (“Ave” → “Avenue”)

• Create link if entities match on
  o EIN or name and any other matching variable
  o Any three matching variables
Matching Results

• Successfully matched 1,608 of 19,550 U.S. parents in 2019 U.S. Direct Investment Abroad data (8.23 percent success rate)
• Entity matches are driven by EIN and company name

Total Matches for Each Variable Among Matched BEA Entities

<table>
<thead>
<tr>
<th>Variable</th>
<th>Match Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>EIN</td>
<td>1,603</td>
</tr>
<tr>
<td>Company Name</td>
<td>1,590</td>
</tr>
<tr>
<td>Url</td>
<td>1,074</td>
</tr>
<tr>
<td>Address</td>
<td>740</td>
</tr>
<tr>
<td>Phone</td>
<td>181</td>
</tr>
</tbody>
</table>
Applications

• Incorporated into auto-editing of 2019 benchmark USDIA survey
• Compustat data was used to help inform the estimates of direct investment earnings for the quarterly surveys during the first quarters of 2020.

• Continuing work to improve the match rate
Spotlight: beaapi
A Python library for BEA’s API
How to improve access to BEA data?

• BEA has an API that allows programmatically collecting data, but data is returned in a way that requires additional work to parse

```json
{"Ordinal":6,"Name":"CL_UNIT","DataType":"string","IsValue":0},
{"Ordinal":7,"Name":"UNIT_MULT","DataType":"numeric","IsValue":0},
{"Ordinal":8,"Name":"DataValue","DataType":"numeric","IsValue":1},"Data":
{"TableID":"2018","SeriesCode":"DPCERX","LineNumber":1,"LineDescription":"Personal consumption expenditures","TimePeriod":"1999Q1","CL_UNIT":"USD","UNIT_MULT":6,"DataValue":7618691,"NoteRef":"2018"},
{"TableID":"2018","SeriesCode":"DPCERX","LineNumber":1,"LineDescription":"Personal consumption expenditures","TimePeriod":"1999Q2","CL_UNIT":"USD","UNIT_MULT":6,"DataValue":7731528,"NoteRef":"2018"},
```

• A package can provide data that is ready to work with

```python
In [15]: beaData = beaapi.get_data(beaKey, 'NIPA', TableName = 'T10205', Year = 'X', Frequency = 'A')
beaData.loc[beaData["LineDescription"] == 'Final sales of computers']
```

```
<table>
<thead>
<tr>
<th>TableName</th>
<th>SeriesCode</th>
<th>LineNumber</th>
<th>LineDescription</th>
<th>TimePeriod</th>
<th>METRIC_NAME</th>
<th>CL_UNIT</th>
<th>UNIT_MULT</th>
<th>DataValue</th>
<th>NoteRef</th>
</tr>
</thead>
<tbody>
<tr>
<td>T10205</td>
<td>BB01RC</td>
<td>17</td>
<td>Final sales of computers</td>
<td>1978</td>
<td>Current Dollars</td>
<td>Level</td>
<td>6</td>
<td>11749</td>
<td>T10205,T10205.3</td>
</tr>
<tr>
<td>T10205</td>
<td>BB01RC</td>
<td>17</td>
<td>Final sales of computers</td>
<td>1979</td>
<td>Current Dollars</td>
<td>Level</td>
<td>6</td>
<td>15721</td>
<td>T10205,T10205.3</td>
</tr>
<tr>
<td>T10205</td>
<td>BB01RC</td>
<td>17</td>
<td>Final sales of computers</td>
<td>1980</td>
<td>Current Dollars</td>
<td>Level</td>
<td>6</td>
<td>20164</td>
<td>T10205,T10205.3</td>
</tr>
</tbody>
</table>
```
• While we have a package for R (bea.R), there is demand for Python:
  o Users are asking for a Python package
  o Python is the dominant DS language

• We are building a Python package beaapi.
What is beaapi?

Two core functions:

1. Return data from the API in a useful structure

2. A faster way to search within select BEA datasets and build API queries.
Search and Explore the API

• Quickly find required API call parameters and search for tables

```python
In [12]: beaapi.get_parameter_list(beaKey, 'NIPA')
Out[12]:
```

<table>
<thead>
<tr>
<th>ParameterName</th>
<th>ParameterDataType</th>
<th>ParameterDescription</th>
<th>ParameterIsRequiredFlag</th>
<th>ParameterDefaultValue</th>
<th>MultipleAcceptedFlag</th>
<th>AllValue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>string</td>
<td>A - Annual, Q-Quarterly, M-Monthly</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ShowMillions</td>
<td>string</td>
<td>A flag indicating that million-dollar data should be used.</td>
<td>0</td>
<td>N</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>TableID</td>
<td>integer</td>
<td>The standard NIPA table identifier</td>
<td>0</td>
<td>&lt;NA&gt;</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>TableName</td>
<td>string</td>
<td>The new NIPA table identifier</td>
<td>0</td>
<td>&lt;NA&gt;</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Year</td>
<td>integer</td>
<td>List of year(s) of data to retrieve (X for All)</td>
<td>1</td>
<td>1</td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>
Analysis with beaapi

- Allows easy filtering and plotting data in Python

Aim to release on GitHub and Python Package index later this year
Spotlight: NowCasting QCEW
Estimating statistics at improved level of detail in Regional

- Often a conflict between timely and detailed (industry) data

<table>
<thead>
<tr>
<th>State Employment</th>
<th>Industry detail (NAICS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annual Estimate</td>
<td>3- / 4-digit</td>
</tr>
<tr>
<td>Quarterly Estimate</td>
<td>2-digit</td>
</tr>
</tbody>
</table>

- Regional would like to extend Quarterly estimates to the same industry detail, but
  - Annual source data (QCEW) is produced with a long lag.
  - Timely data (CES) used for leading quarter, has good coverage only at the 2-digit level. At 3-/4-digit level it has many missing values.
Good coverage for larger states and specialty industries
• Project goal is to nowcast missing CES data
  o If successful could also help improve the detail for Regional Personal Income and GDP

• Difficult as imputations needs to satisfy both geographic and industry hierarchical constraints

• Similar to challenges with other BEA statistics, so evaluate existing in-house tools (KRAS)
• KRAS is a generalized iterative scaling method that can reconcile large systems of linear equations under conflicting external information and inconsistent constraints.
  o Appropriate for hierarchical data (geographies, industries).
  o Resolving conflicting information is an important feature because CES National and CES State data are not consistent.
  o Already in use by NEA to balance Input-Output tables.

• The proposed method utilizes additional CES industry detail in a KRAS framework to impute missing information.

• KRAS results are currently being compared to a naive estimation approach (autoregressive model with quarterly dummy variables).
Appendix
Data Science: Definition
Innovation at BEA

Data Science Strategy

Communications Training
Pilot projects
Other cross-cutting areas

Python Implementation
Guiding Principles

Principle 1. **BEA is primed for a data science strategy.**

Principle 2. **Data science is just one piece of the data pie.**

Principle 3. **BEA’s data science strategy is “living.”**

Principle 4. **Data science is a journey—with a destination.**

Principle 5. **Data science doesn’t work in a vacuum.**
1. Adopt a bureau-wide data science definition that (1) centers around a data science process with three inter-related parts (extract-transform-load, analyze, and communicate) and (2) emphasizes the importance of operational forces, including training, IT resources, hiring and partnerships, and customer service.

2. Take a hybrid approach to carrying out data science at BEA, with roles and responsibilities shared by an executive-level office, a Data Science Coordinating Team, and existing cross-directorate groups.

3. Establish a bureau-wide IT posture that readily harnesses cutting-edge data science technologies while (1) maintaining security controls and (2) efficiently using government resources.

4. Stand up a comprehensive Data Science Training Program that provides multiple avenues for staff to engage with cutting-edge tools, techniques, and technologies.

5. Strengthen BEA’s data science knowledge base through strategic hiring and partnerships across BEA program areas and with other agencies, the private sector, and academia.

6. Prioritize customer service throughout the data science process.
1. In every BEA program area, Python should be adopted as the language of choice for all code involved in the production of BEA statistical products.

2. Two applications should be made available to staff for research or analytical activities not involving production code: SAS and R.

3. Each program area should determine a schedule detailing when it will be practical to cease all new desktop code development in six applications: Access, BPRL, FAME, MATLAB, Stata, and SQL.
## DSCT Action Items

<table>
<thead>
<tr>
<th>Short-term progress (Past 6 months)</th>
<th>Medium-term plans (Ongoing/Year 1)</th>
<th>Long-term goal (2–5 years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Map out data science roles and responsibilities for cross-directorate groups</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Establish coordination strategy with DAIT</td>
<td>• Meet regularly with DAIT to identify joint efforts</td>
<td>• Continue cross-group coordination</td>
</tr>
<tr>
<td>• Speak on strategy in Research Town Hall</td>
<td>• Co-host Data Science Open House</td>
<td>• Support bureau-wide communication on data science</td>
</tr>
<tr>
<td>Begin work on developing a BEA Data Science Curriculum and Toolkit</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Develop draft personas and pilot strategy, working with DAIT</td>
<td>• Continue developing training strategy with DAIT</td>
<td>• Develop robust curriculum</td>
</tr>
<tr>
<td>• Sponsor Machine Learning for Economists training</td>
<td>• Conduct pilot test with at least one volunteer per persona</td>
<td>• Feature/sponsor trainings</td>
</tr>
<tr>
<td>Catalog connections between BEA’s data science strategy and outside stakeholders</td>
<td>• Establish communities of practice</td>
<td>• Gather feedback and improve curriculum</td>
</tr>
<tr>
<td>• Develop explainer and post to Data Science Intranet page</td>
<td>• Develop and update feedback mechanisms (e.g., intranet page, Bureau Beat, meetings with BEA reps to outside groups)</td>
<td>• Continue coordination with BEA reps to outside groups</td>
</tr>
<tr>
<td>Develop a BEA Data Science Project Model as a proof of concept</td>
<td>• Explore pilot projects with DAIT</td>
<td>• Support bureau-wide communication on these efforts</td>
</tr>
<tr>
<td>• Explore project management structures, including Agile/DevOps</td>
<td>• Sponsor pilot projects</td>
<td></td>
</tr>
<tr>
<td>Design the infrastructure for tracking data science projects</td>
<td>• Test processes and gather lessons learned</td>
<td>• Continue pilots</td>
</tr>
<tr>
<td>• Develop major projects list</td>
<td></td>
<td>• Apply lessons learned to develop best practices</td>
</tr>
<tr>
<td>• Define “big data” projects</td>
<td>• Catalog “big data” projects</td>
<td></td>
</tr>
<tr>
<td>• Develop project tracking method</td>
<td>• Feature and update project list on intranet</td>
<td>• Deploy project tracking method</td>
</tr>
<tr>
<td>• Spotlight cross-cutting projects that feature advanced techniques</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>