Data Science @ BEA

Brian Quistorff, Data Science Coordinating Team

Amanda Lyndaker, Desktop Applications Implementation Team

BEA Advisory Committee Meeting: November 18, 2022

Bureau of Economic Analysis
U.S. DEPARTMENT OF COMMERCE

Session Overview

- Update on modernizing BEA's Data Science strategy and implementation

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

- Comments from Discussants

- Q&A

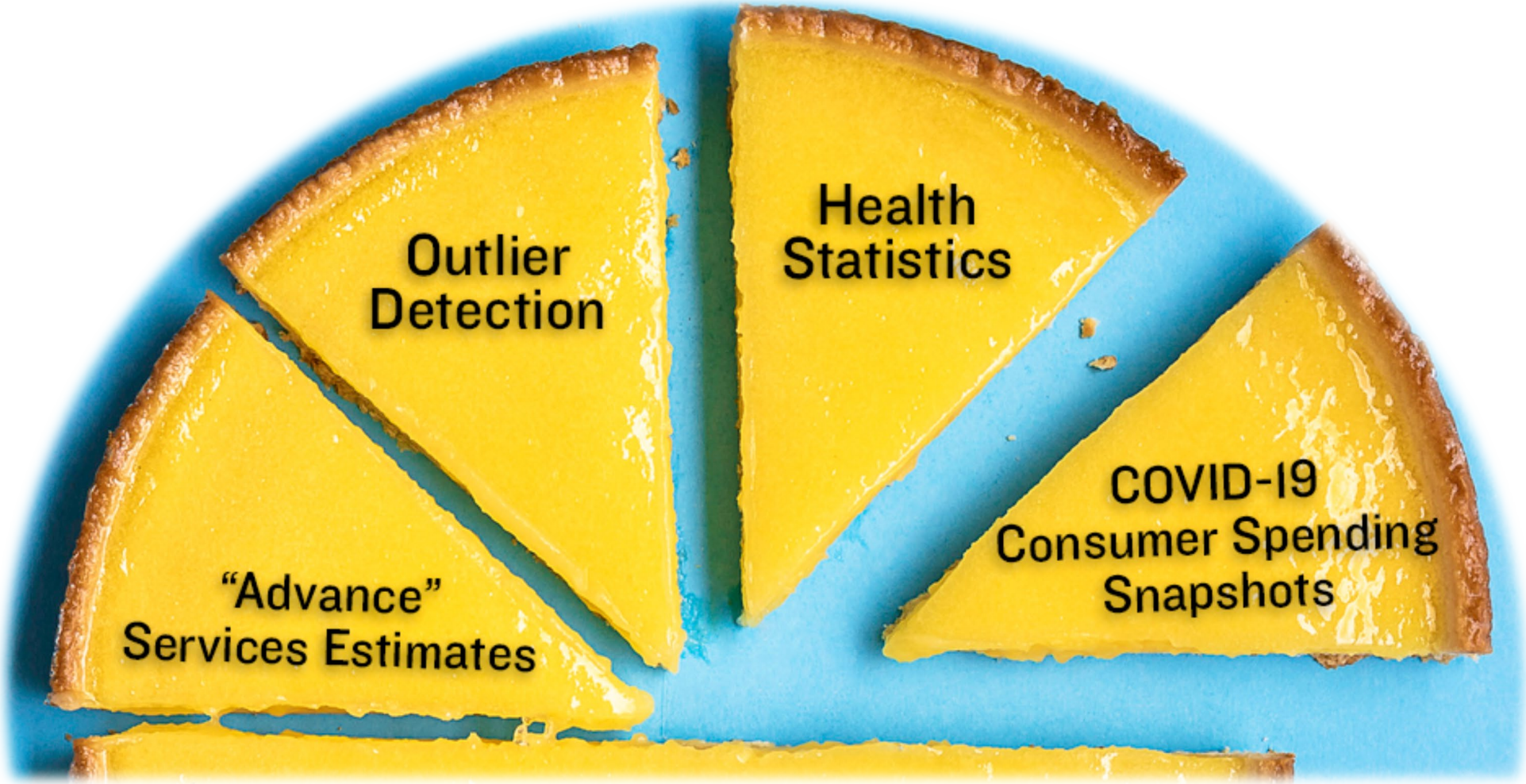
Data Science Vision

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

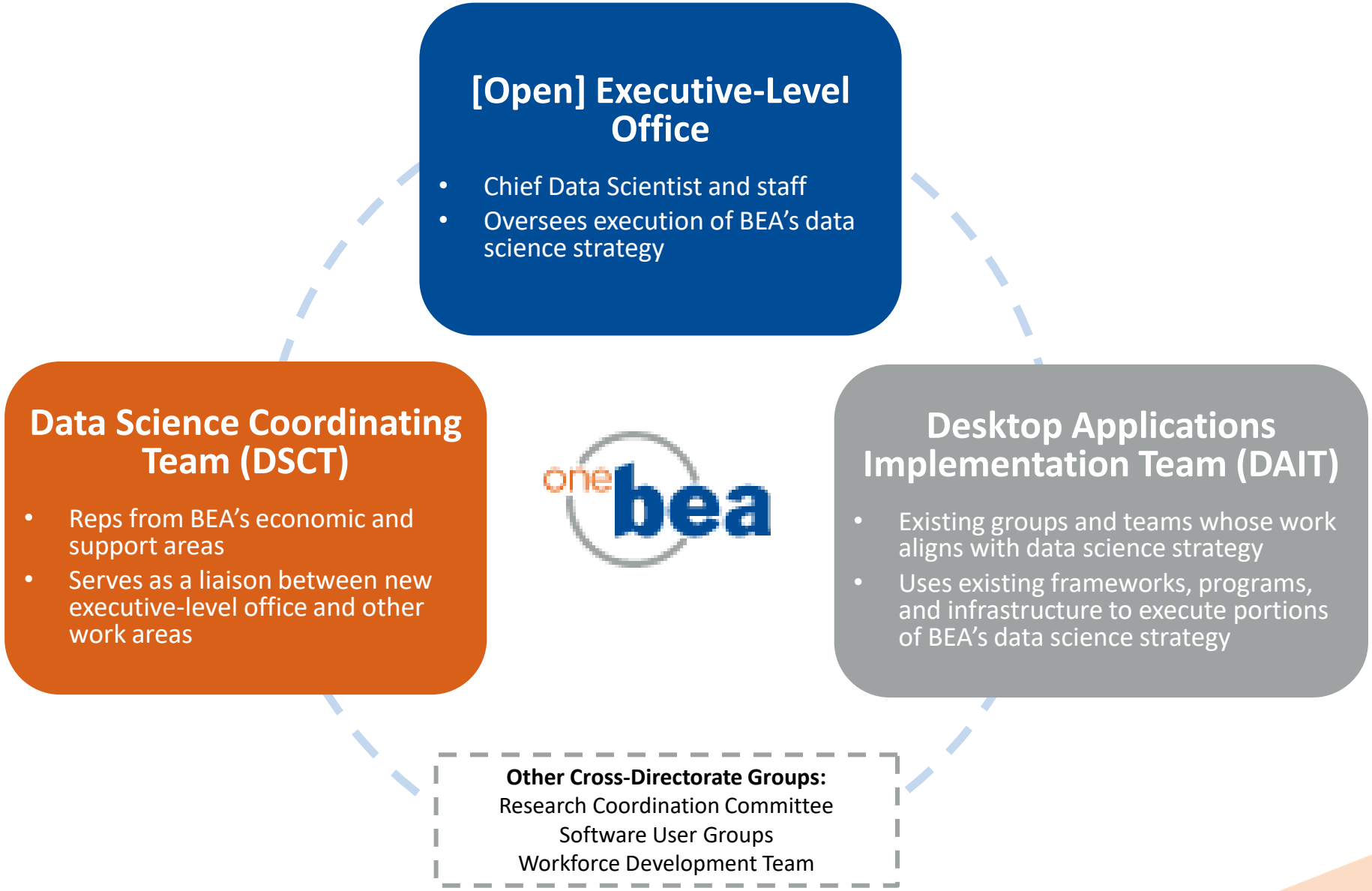
Mission goals: Data science touches on each of BEA's four mission goals—(1) accuracy and reliability, (2) relevance, (3) customer service, (4) operational excellence

Data Science Uses at BEA

Data Science: Slices of Success



Data Science: Operational Framework



DSCT Priority Highlights

- Developing a BEA Data Science Curriculum and Toolkit
  - Develop draft personas and pilot strategy, working with DAIT
  - Sponsor Machine Learning for Economists training
  - Continue developing training strategy with DAIT

- Developing a BEA Data Science Project Model as a proof of concept
  - Explore project management structures, including Git and Agile/DevOps
  - 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

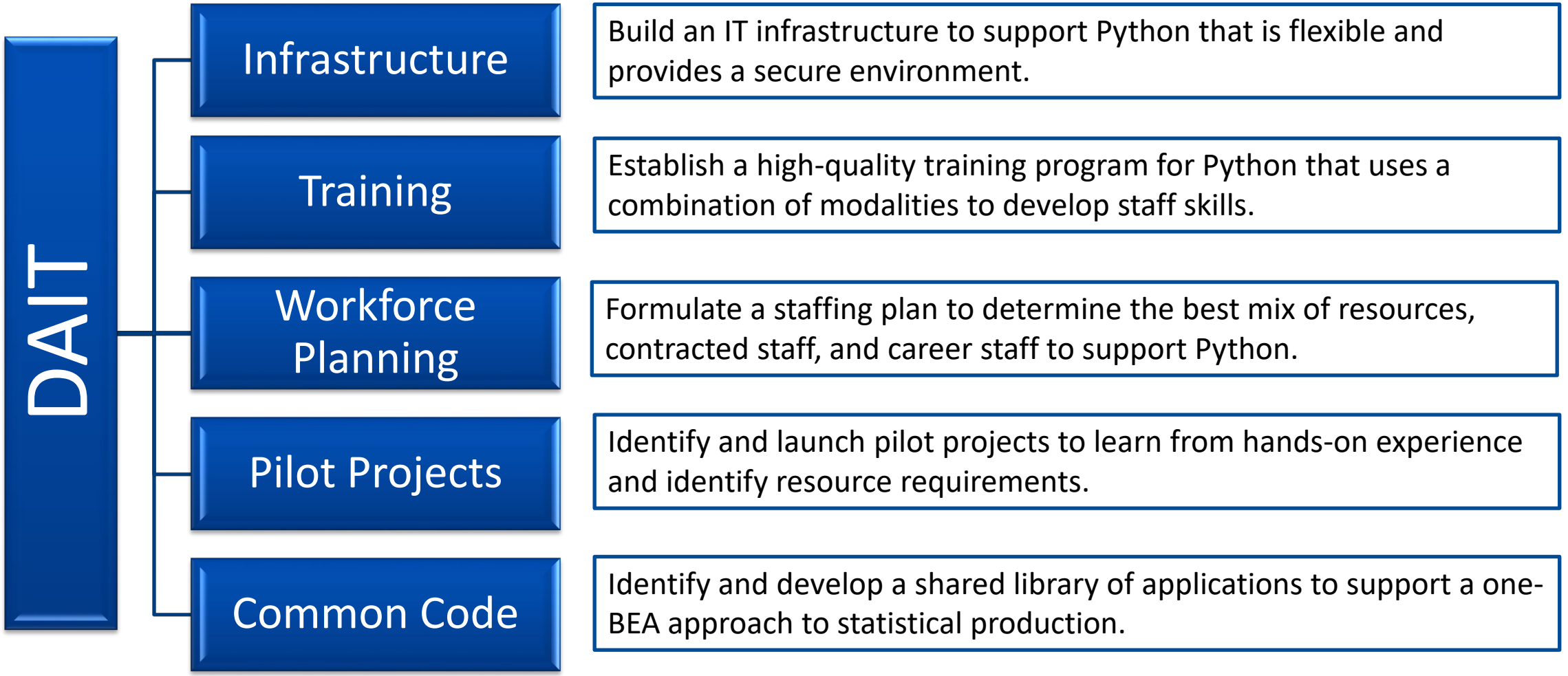
Process

- 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



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

Larkin Terrie

Bureau of Economic Analysis
U.S. DEPARTMENT OF COMMERCE

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
  - 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
  - Drop special characters (&-,.%/#)
  - From company names, drop common words and abbreviations (“Inc.”, “Limited”)
  - In address, standardize common abbreviations (“Ave” → “Avenue”)
- Create link if entities match on
  - EIN or name and any other matching variable
  - 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

Variable	Match Count
EIN	1,603
Company Name	1,590
Url	1,074
Address	740
Phone	181

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

Andrea Batch

Bureau of Economic Analysis
U.S. DEPARTMENT OF COMMERCE

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


```
{"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":"7,618,691","NoteRef":"2018"},
{"TableID":"2018","SeriesCode":"DPCERX","LineNumber":"1","LineDescription":"Personal consumption
expenditures","TimePeriod":"1999Q2","CL_UNIT":"USD","UNIT_MULT":"6","DataValue":"7,731,528","NoteRef":"2018"},
```


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

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

Out[15]:

	TableName	SeriesCode	LineNumber	LineDescription	TimePeriod	METRIC_NAME	CL_UNIT	UNIT_MULT	DataValue	NoteRef
1412	T10205	BB01RC	17	Final sales of computers	1978	Current Dollars	Level	6	11749	T10205,T10205.3
1413	T10205	BB01RC	17	Final sales of computers	1979	Current Dollars	Level	6	15721	T10205,T10205.3
1414	T10205	BB01RC	17	Final sales of computers	1980	Current Dollars	Level	6	20164	T10205,T10205.3

- While we have a package for R (bea.R), there is demand for Python:
  - Users are asking for a Python package
  - 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

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

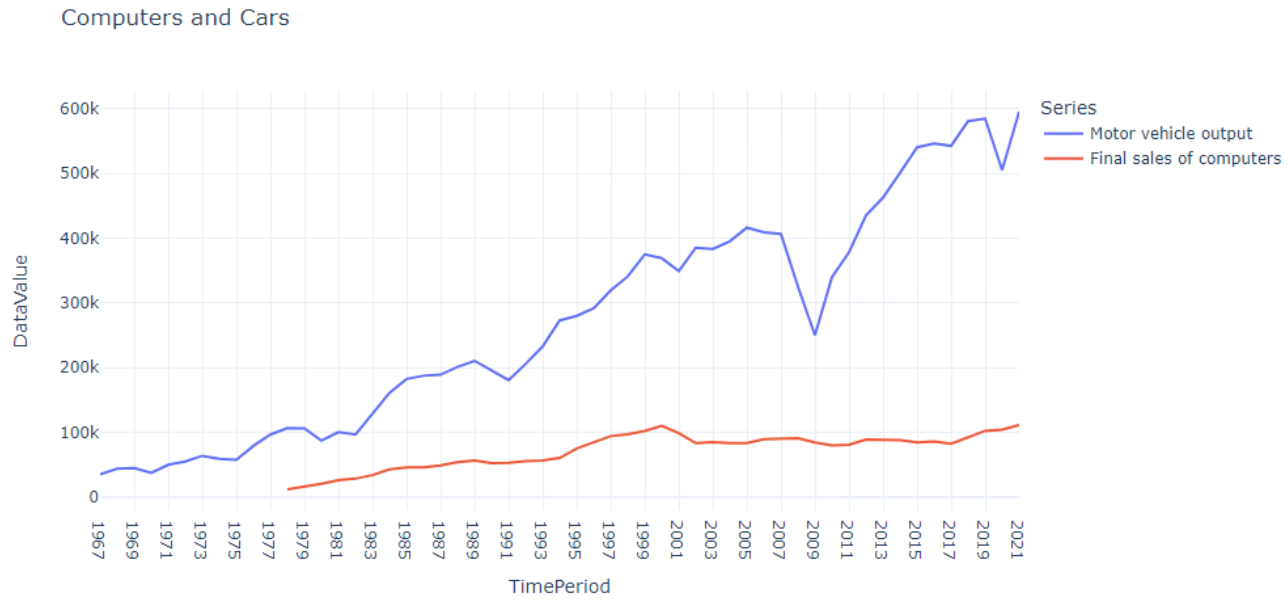
Out[12]:

	ParameterName	ParameterDataType	ParameterDescription	ParameterIsRequiredFlag	ParameterDefaultValue	MultipleAcceptedFlag	AllValue
0	Frequency	string	A - Annual, Q-Quarterly, M-Monthly	1		1	
1	ShowMillions	string	A flag indicating that million-dollar data sho...	0	N	0	
2	TableID	integer	The standard NIPA table identifier	0	<NA>	0	
3	TableName	string	The new NIPA table identifier	0	<NA>	0	
4	Year	integer	List of year(s) of data to retrieve (X for All)	1		1	X

Analysis with beaapi

- Allows easy filtering and plotting data in Python

```
In [19]: cars_and_comps = beaData.loc[((beaData['LineDescription'] == 'Final sales of computers') | (beaData['LineDescription'] == 'Motor
beaplot = px.line(cars_and_comps, x=cars_and_comps['TimePeriod'], y=cars_and_comps['DataValue'], width=1000, height=500,
                  labels={
                     "x": cars_and_comps['TimePeriod']
                  }, title=f'Computers and Cars', color=cars_and_comps['LineDescription'], template='plotly_white')
beaplot.update_layout(legend_title="Series")
beaplot
```



Aim to release on GitHub and Python Package index later this year

Spotlight: NowCasting QCEW

Gabriel Medeiros

Bureau of Economic Analysis
U.S. DEPARTMENT OF COMMERCE

Estimating statistics at improved level of detail in Regional

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

State Employment	Industry detail (NAICS)
Annual Estimate	3- / 4-digit
Quarterly Estimate	2-digit

- 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-digitl level. At 3-/4-digit level it has many missing values.

CES Data Availability (snapshot)

States

Naics-3/4

= Coverage

Good coverage for larger states and specialty industries

Nowcasting QCEW

- Project goal is to nowcast missing CES data
  - 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

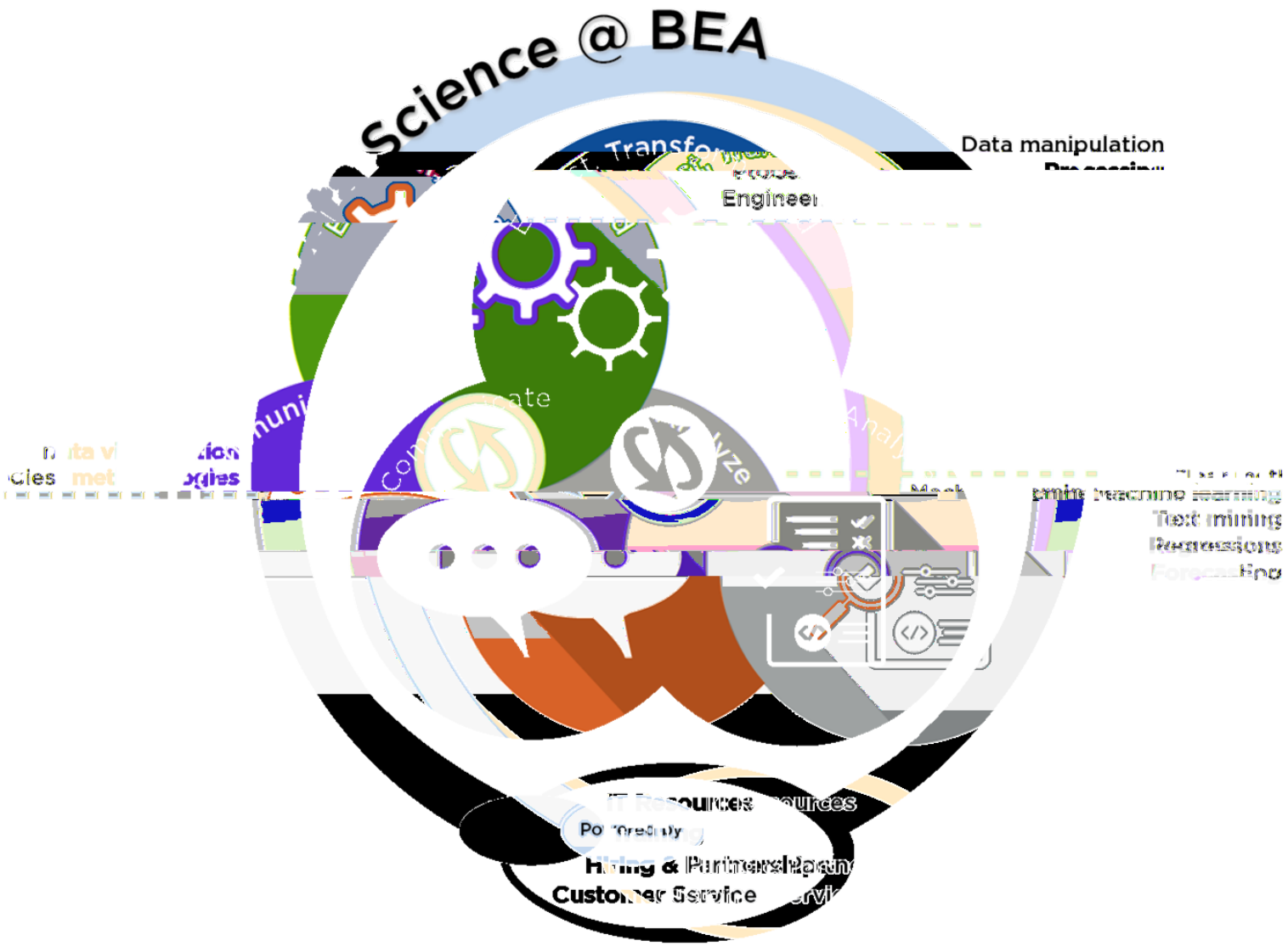
- Similar to challenges with other BEA statistics, so evaluate existing in-house tools (KRAS)

KRAS

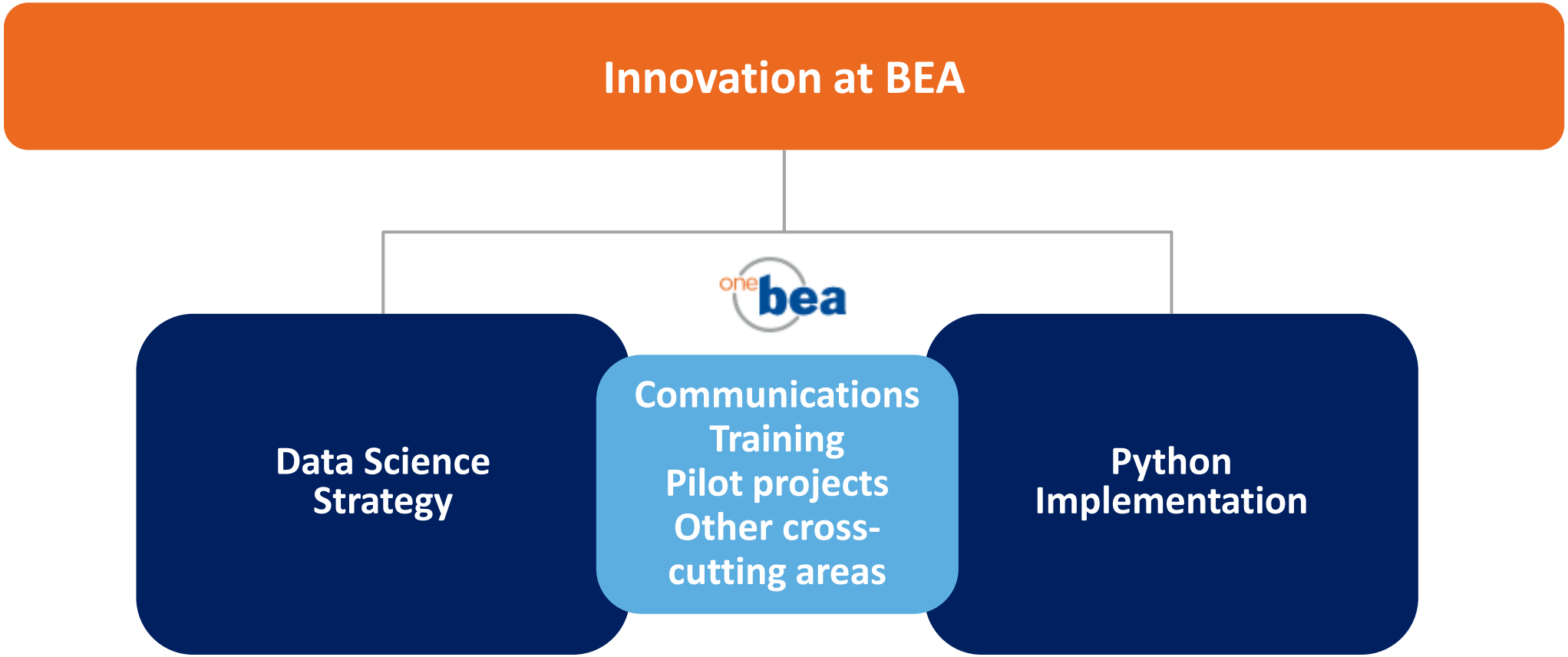
- KRAS is a generalized iterative scaling method that can reconcile large systems of linear equations under conflicting external information and inconsistent constraints.
  - Appropriate for hierarchical data (geographies, industries).
  - Resolving conflicting information is an important feature because CES National and CES State data are not consistent.
  - 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



DAIT and DSCT



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.**

Data Science Strategy Working Group Recommendations

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.

Desktop Applications Assessment Working Group Recommendations

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

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