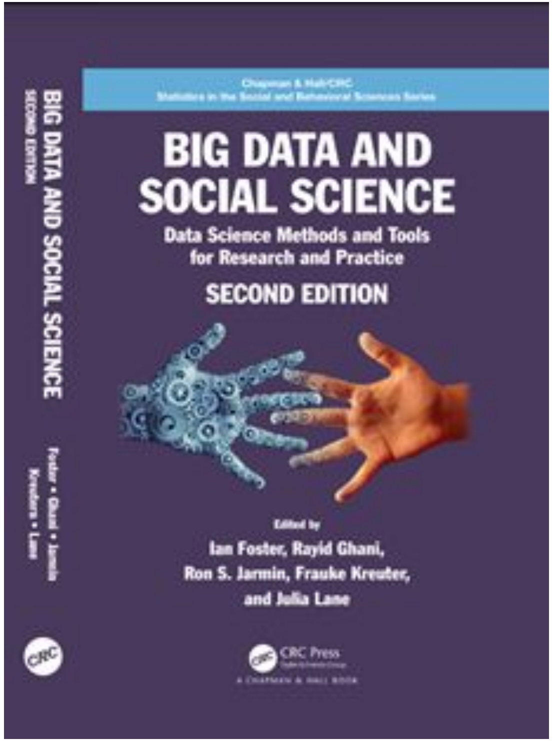
Enhancing Survey Data with Public Data and Text Analysis

Benjamin Feder
Julia Lane
Ekaterina Levitskaya
Clayton Hunter



Chapman & Hall/CRC
Statistics in the Social and Behavioral Sciences Series

BIG DATA AND SOCIAL SCIENCE

Data Science Methods and Tools for Research and Practice

SECOND EDITION

Edited by
Ian Foster, Rayid Ghani,
Ron S. Jarmin, Frauke Kreuter,
and Julia Lane

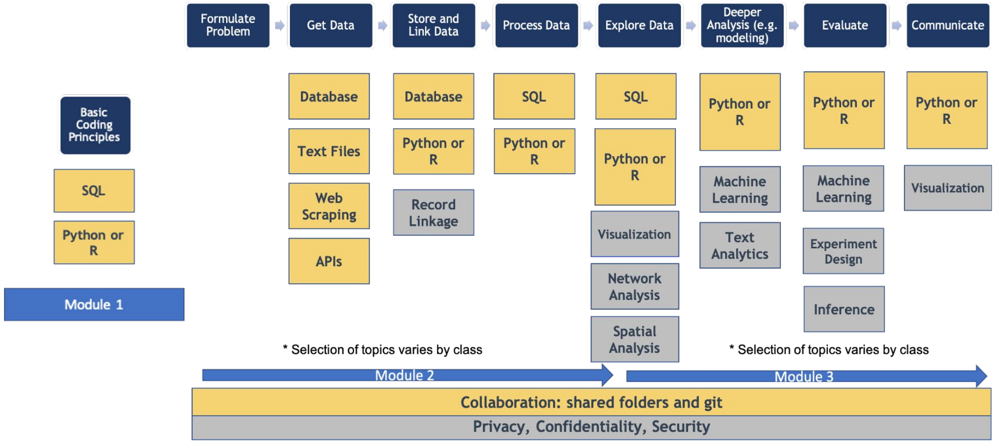
CRC Press
A CHAPMAN & HALL BOOK

BIG DATA AND SOCIAL SCIENCE
SECOND EDITION
Foster • Ghani • Jarmin
Kreuter • Lane

With many thanks to Dr. Brian Kim, Institute for Research on Innovation and Science, National Science Foundation, the Coleridge Initiative team, and past participants of the Applied Data Analytics training programs

The Goal of the Applied Data Analytics Training Classes

- Build the technical and analytical capacity of public sector employees to work with real-world, confidential data;
- Demonstrate the value of working with such data in a secure FedRAMP environment across jurisdictional (state and agency) lines;
- Create a collaborative community of researchers and practitioners within and beyond the program.



Basic Coding Principles	Formulate Problem	Get Data	Store and Link Data	Process Data	Explore Data	Deeper Analysis (e.g. modeling)	Evaluate	Communicate
SQL		Database	Database	SQL	SQL	Python or R	Python or R	Python or R
Python or R		Text Files	Python or R	Python or R	Python or R	Machine Learning	Machine Learning	Visualization
		Web Scraping	Record Linkage		Visualization	Text Analytics	Experiment Design	
		APIs			Network Analysis		Inference	
					Spatial Analysis			

Module 1

* Selection of topics varies by class

* Selection of topics varies by class

Module 2

Module 3

Collaboration: shared folders and git

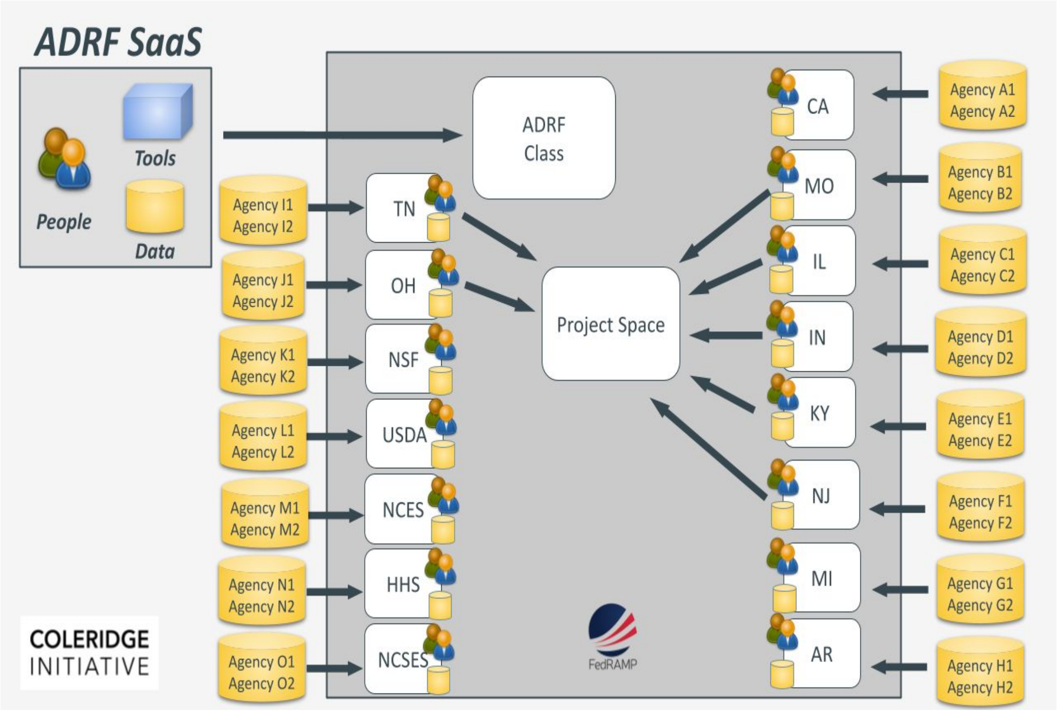
Privacy, Confidentiality, Security

What is the value?

1. Multi-agency innovation sandbox to develop new products

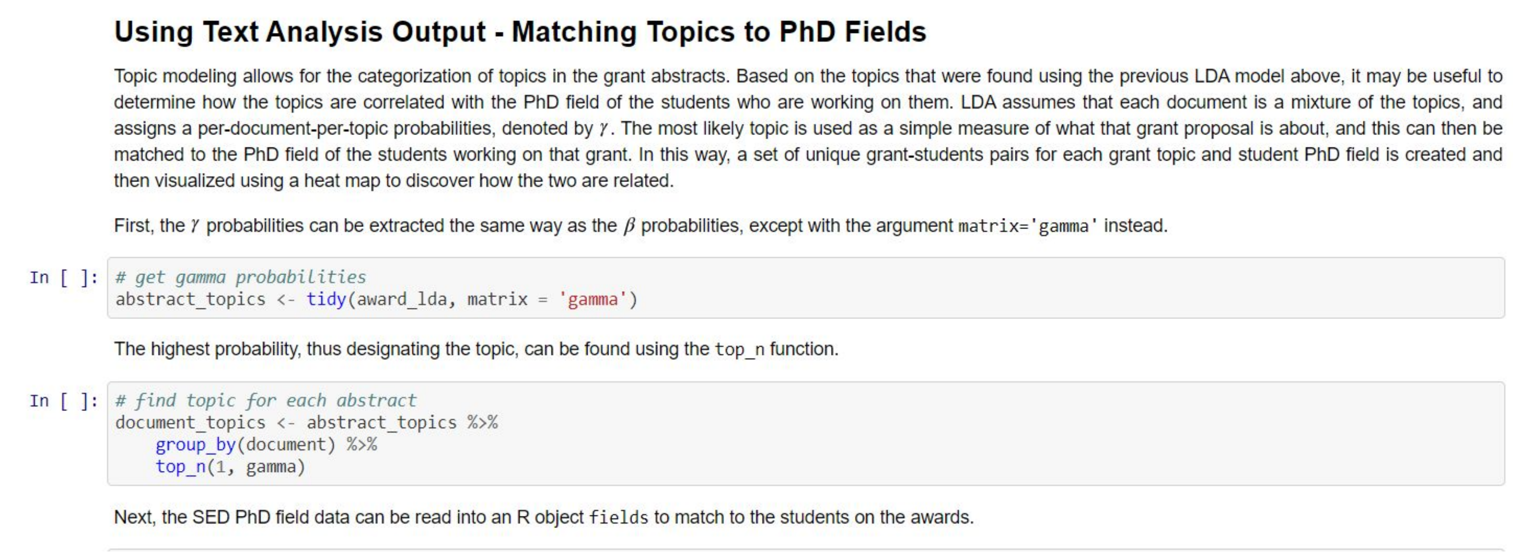
2. Improved ability of government agencies to use their own data for evidence-based policymaking

3. New networks across state and agency lines



How is the value created?

- Lectures and hands-on, group-based working sessions
- Interactive Jupyter notebooks
- Final projects



Using Text Analysis Output - Matching Topics to PhD Fields

Topic modeling allows for the categorization of topics in the grant abstracts. Based on the topics that were found using the previous LDA model above, it may be useful to determine how the topics are correlated with the PhD field of the students who are working on them. LDA assumes that each document is a mixture of the topics, and assigns a per-document-per-topic probabilities, denoted by γ. The most likely topic is used as a simple measure of what that grant proposal is about, and this can then be matched to the PhD field of the students working on that grant. In this way, a set of unique grant-students pairs for each grant topic and student PhD field is created and then visualized using a heat map to discover how the two are related.

First, the γ probabilities can be extracted the same way as the β probabilities, except with the argument `matrix='gamma'` instead.

```
In [ ]: # get gamma probabilities
        abstract_topics <- tidy(award_lda, matrix = 'gamma')
```

The highest probability, thus designating the topic, can be found using the `top_n` function.

```
In [ ]: # find topic for each abstract
        document_topics <- abstract_topics %>%
            group_by(document) %>%
            top_n(1, gamma)
```

Next, the SED PhD field data can be read into an R object `fields` to match to the students on the awards.

What is the evidence of success?

- New linked data assets
- Trained workforce
- New products and ideas for agencies to address key challenges
    - Multi-State Post-Secondary Dashboard
    - Unemployment to Reemployment Portal

Image by Benjamin Feder, Coleridge Initiative

Contact Information

Benjamin Feder

Research Statistician

ben.feder@coleridgeinitiative.org

Julia Lane

Co-Founder and Director

julia.lane@coleridgeinitiative.org

Ekaterina Levitskaya

Associate Research Scientist

ekaterina.levitskaya@coleridgeinitiative.org

Enhancing Survey Data with Public Data and Text Analysis

Benjamin Feder
Julia Lane
Ekaterina Levitskaya
Clayton Hunter

With many thanks to Dr. Brian Kim, Institute for Research on Innovation and Science, National Science Foundation, the Coleridge Initiative team, and past participants of the Applied Data Analytics training programs.

Table of Contents

Text Analysis Jupyter Notebook

The jupyter notebook provides a walkthrough on how text analysis methods can be used to parse grant abstracts from Federal RePORTER into specific topics of research to better understand doctoral recipients' academic careers. The notebook is part of a series of learning materials created for the Fall 2020 Coleridge Initiative Applied Data Analytics training program hosted by the National Center of Science and Engineering Statistics. This particular notebook can be accessed through this link.

Coleridge Initiative Training Website

The Coleridge Initiative's Applied Data Analytics programs are designed to train government employees and public policy analysts how to tackle important policy problems. For more information, visit https://coleridgeinitiative.org/training/.

Change Through Data: A Data Analytics Training Program for Government Employees

Frauke Kreuter, Rayid Ghani, and Julia Lane published an article in the Harvard Data Science Review detailing the vision behind the Applied Data Analytics training programs. The article, *Change Through Data: A Data Analytics Training Program for Government Employees*, can be accessed here.

In-Demand Questions Addressed by Training Programs

Team projects have addressed various types of in-demand questions. Three examples, reemploying the unemployed, reducing recidivism with employment, and returns on investment in higher education, are highlighted in the following sections.

Coleridge Initiative
Applied Data Analytics Training

COLERIDGE
INITIATIVE

OPENING OPPORTUNITIES FOR STATES TO LEARN & COLLABORATE

FOCUS AREA: Reemploying the Unemployed

COVID-19 devastated the U.S. labor market with unequal impacts across groups, regions, and industries. Given the sudden nature of the crisis, Unemployment Insurance (UI) claims shattered records. The weekly number of individuals receiving benefits rose exponentially from 2 million in early March to 30 million in June, putting extreme pressure on states' UI trust funds. A significant number of these individuals have permanently lost their jobs and will require additional education and training to find new positions and succeed in the new economy. Thus, assisting the unemployed in returning to work is pivotal to manage UI duration and UI trust fund outflows.

It has become vital that state and local agencies use their data to address these reemployment challenges. Local workforce leaders require timely and relevant information on characteristics of UI claimants, dynamics of UI receipt, and employment opportunities within local industries. This information is useful in effectively counseling job seekers and strategically targeting scarce public training funds.

Through the Applied Data Analytics training program, states in the Midwest Collaborative began to harness the potential of their weekly UI claims data by creating a powerful, new reemployment tool, the Unemployment to Reemployment Dashboard. It visualizes data on new and continuing UI claims by week, geographical location, industry, and the characteristics of the UI claimants. The Dashboard encompasses data of labor markets spanning multiple states, thanks to the data sharing agreements between states. As the economic cycle progresses, states will add training providers, RESEA, TANF, and other data to derive insights from effective reemployment achieved through program participation.

THE FUTURE IS NOW *for states to generate ideas and products focused on the current jobs crisis and other policy concerns. Consider the opportunities provided by the Coleridge Initiative's Applied Data Analytics program.*

FOR MORE INFORMATION: **coleridgeinitiative.org/training/**

COLERIDGE
INITIATIVE

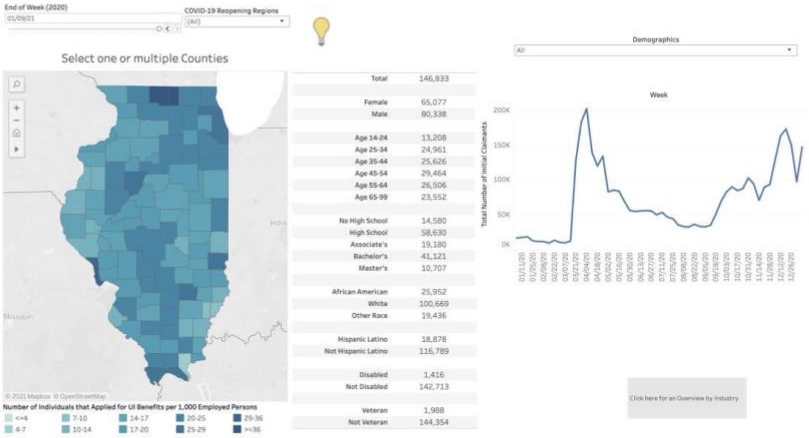
How did the Applied Data Analytics training program help?

1. *Trained State Agency Staff*

The Coleridge Initiative's Applied Data Analytics program provided participants with **hands-on training in modern data analytics**. The curriculum covered data management, data analysis, machine learning, inference, privacy protection, and the ethical use of data. Participants used real-world data focused on a just-in-time policy concern: reemploying the unemployed. State teams included both technical and program staff, which helped bridge knowledge gaps across disciplines and strengthened relationships.

2. *Enabled States to Create New Products*

The Applied Data Analytics classes are an **"innovation sandbox."** Participants explored ideas and analytical techniques with colleagues from within their state as well as colleagues from across the country. This rich network of individuals and teams generated new research ideas with a common research focus. The result is the ***Unemployment to Reemployment Dashboard***, an innovative, multi-state data product.



End of Week (2020): 01/09/21

COVID-19 Reopening Regions: (All)

Select one or multiple Counties

Total	146,833
Female	65,077
Male	80,338
Age 14-24	13,208
Age 25-34	24,961
Age 35-44	25,626
Age 45-54	29,464
Age 55-64	26,506
Age 65-99	23,552
No High School	14,580
High School	58,630
Associate's	19,180
Bachelor's	41,121
Master's	10,707
African American	25,952
White	100,669
Other Race	19,436
Hispanic Latino	18,878
Not Hispanic Latino	116,789
Disabled	1,416
Not Disabled	142,713
Veteran	1,988
Not Veteran	144,354

Demographics: All

3. *Provided a Secure Place for States to Share Data*

The Coleridge Initiative's Administrative Data Research Facility (ADRF) enabled state agencies to **share real-world data securely and safely**. During training, state agency staff learned how this federally-certified, secure platform applies a "five safes" framework – safe project, safe people, safe settings, safe data and safe output – to ensure data are protected.

"This environment awakens you to all the datasets, especially those that can be linked across states. And because states have similar policy questions, we can reduce duplicative work. We can share code and methodologies and get to studies and outcomes faster."

- State LMI Director, March 2019

Over 500 people from more than 100 state agencies successfully completed an Applied Data Analytics program over the past three years. Together they initiated over 100 exploratory research projects focusing on critical policy concerns.

"The analytic possibilities presented by the ADRF have great potential to improve our understanding of fundamental issues facing our state. It could lead to policy decisions that are much more informed than has previously been possible."

- State LMI Director, August 2020

FOR MORE INFORMATION: **coleridgeinitiative.org/training/**

Reducing Recidivism with Employment

Coleridge Initiative
Applied Data Analytics Training

OPENING OPPORTUNITIES FOR STATES TO LEARN & COLLABORATE

FOCUS AREA: Reducing Recidivism with Employment

The United States has the highest incarceration rate in the world, and with over two million Americans currently incarcerated, the United States accounts for roughly one-quarter of the world's incarcerated population. The financial cost of incarceration to society is roughly $180 billion annually after factoring in judicial, legal, and policing costs. The social cost is also of concern as the incarceration rate is substantially higher for Americans of color. One of the major contributors to these staggering numbers is recidivism. Fifty percent of individuals return to prisons or jails within a few years of their release, and the racial discrepancies concerning incarceration and recidivism make addressing recidivism an even greater public priority.

State and local agencies and their strategic partners need timely information and evidence to effectively employ interventions intended to reduce recidivism rates and help individuals succeed in the labor market after their release. Thanks to the Applied Data Analytics program, several state teams worked together to obtain meaningful insights on post-incarceration employment experiences and the impact of recidivism reduction efforts by linking corrections and workforce data. Using data from one state, the teams generated novel project ideas and initiated a new research practice of cumulative learning with the potential to expand nationally.

Examples of questions they posed to the data include:

- What are the most common employment patterns of prisoners following their release?
- How does employment, wage levels, gang involvement, substance abuse, educational attainment, or children's presence affect the likelihood of recidivism?
- How do prison programs that support or address educational attainment, social skills, and therapy affect the likelihood of recidivism?
- How do predictive tools identify individuals more likely to recidivate, and how can we use this information to help design programs that will provide enhanced support to these individuals?

***THE FUTURE IS NOW** for states to generate ideas and products focused on reducing recidivism rates.*
Consider the opportunities provided by the Coleridge Initiative's Applied Data Analytics program.

FOR MORE INFORMATION: **coleridgeinitiative.org/training/**

COLERIDGE
INITIATIVE

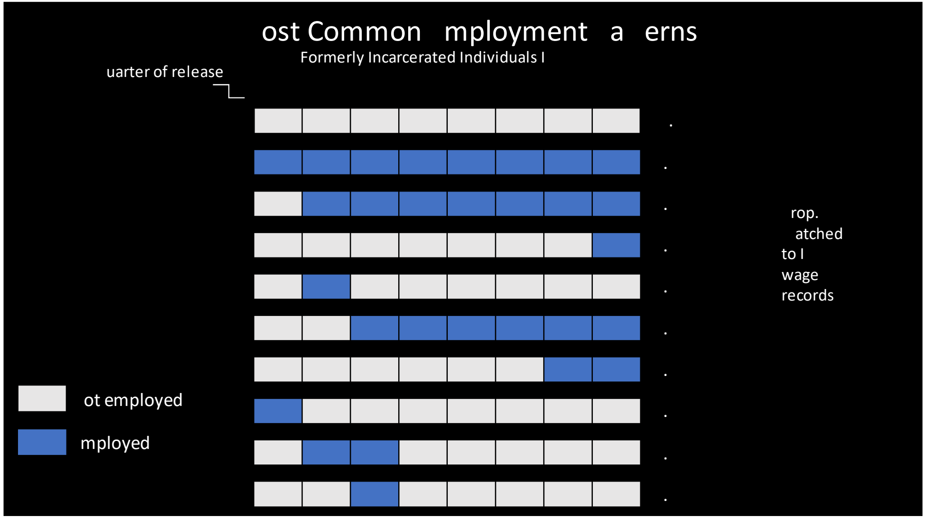
How did the Applied Data Analytics training program help?

1. *Trained State Agency Staff*

The Coleridge Initiative's Applied Data Analytics program provided participants with **hands-on training in modern data analytics**. The curriculum covered data management, data analysis, machine learning, inference, privacy protection, and the ethical use of data. Participants used real-world data focused on a just-in-time policy concern: reducing recidivism. State teams included both technical and program staff, which helped bridge knowledge gaps across disciplines and strengthened relationships.

2. *Enabled States to Create New Products*

The Applied Data Analytics classes are an **"innovation sandbox."** Participants explored ideas and analytical techniques with colleagues from within their state as well as colleagues from across the country. This rich network of individuals and teams generated new research ideas with a common research focus.



3. *Provided a Secure Place for States to Share Data*

The Coleridge Initiative's Administrative Data Research Facility (ADRF) enabled state agencies to **share real-world data securely and safely**. During training, state agency staff learned how this federally-certified, secure platform applies a "five safes" framework – safe project, safe people, safe settings, safe data and safe output – to ensure data are protected.

"Cross state administrative data from the Coleridge training provided timely and relevant information for employers in the Midwest, allowing us to make a case for recruiting new firms to hire in the state."

- State Economic Development Official, November 2020

Over 500 people from more than 100 state agencies successfully completed an Applied Data Analytics program over the past three years. Together they initiated over 100 exploratory research projects focusing on critical policy concerns.

"Our data is the greatest untapped resource for improving public policy outcomes."

- State Workforce Director, August 2020

FOR MORE INFORMATION: **coleridgeinitiative.org/training/**

Returns on Investments in Higher Education

Coleridge Initiative
Applied Data Analytics Training

COLERIDGE
INITIATIVE

OPENING OPPORTUNITIES FOR STATES TO LEARN & COLLABORATE

FOCUS AREA: Returns on Investments in Higher Education

Jobs today require more education and training. Indeed, a greater percentage of new hires have at least some college education than ever before. However, matching the education and training that workers receive with the demand for their skills across the economy requires a new data infrastructure. Despite investing over $80 billion in higher education, states do not have access to timely information about the employment outcomes of those who attended their post-secondary institutions to guide their investments and policies toward higher returns.

Capturing timely and truly representative information on post-secondary employment outcomes has been an elusive goal for state policymakers because of two critical needs. First, **states need to link education data with workforce data**. Second, since many post-secondary graduates move or work out-of-state after graduating, **states need to link their data to the education and workforce data of other states**, not just their own.

Thanks to the Applied Data Analytics training program, several states are demonstrating their ability to respond to these critical information needs. As the Multi-State Postsecondary Partnership, they have created a powerful and visually engaging Multi-State Post-Secondary Feedback Report. The Report provides information to state education leaders and policymakers on critical outcomes information. The report allows them to trace the pathways of their graduates and answer key questions:

- What are the employment and earnings patterns of our graduates, based on major, credential level, institution, and origin (in-state, out-of-state)?
- What do employment patterns suggest about industries or occupations in the highest demand?
- How many graduates are working out-of-state and how do out-of-state wages compare to in-state wages? How do these post-secondary outcomes differ by major and credential level?
- How many graduates who accept jobs in other states are from certain high-demand or essential fields? How does this impact our state workforce and economic development plans?

***THE FUTURE IS NOW** for states to generate ideas and products focused on the returns on investment in higher education. Consider the opportunities provided by the Coleridge Initiative's Applied Data Analytics program.*

FOR MORE INFORMATION: **coleridgeinitiative.org/training/**

COLERIDGE
INITIATIVE

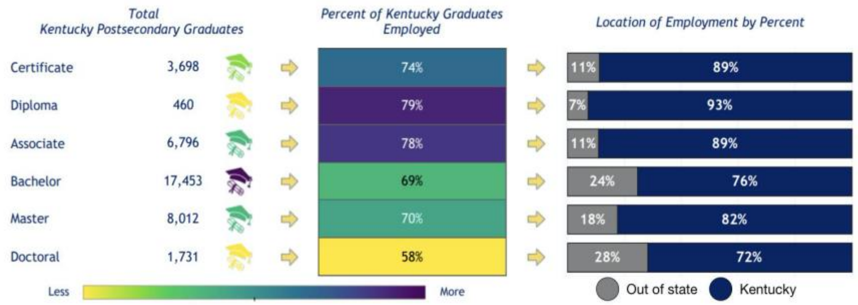
How did the Applied Data Analytics training program help?

1. *Trained State Agency Staff*

The Coleridge Initiative's Applied Data Analytics program provided participants with **hands-on training in modern data analytics**. The curriculum covered data management, data analysis, machine learning, inference, privacy protection, and the ethical use of data. Participants used real-world data focused on a just-in-time policy concern: post-secondary outcomes. State teams included both technical and program staff, which helped bridge knowledge gaps across disciplines and strengthened relationships.

2. *Enabled States to Create New Products*

The Applied Data Analytics classes are an **"innovation sandbox."** Participants explored ideas and analytical techniques with colleagues from within their state as well as colleagues from across the country. This rich network of individuals and teams generated new research ideas with a common research focus. The result is the ***Multi-State Post-Secondary Feedback Report***, an innovative, multi-state data product.



3. *Provided a Secure Place for States to Share Data*

The Coleridge Initiative's Administrative Data Research Facility (ADRF) enabled state agencies to **share real-world data securely and safely**. During training, state agency staff learned how this federally-certified, secure platform applies a "five safes" framework – safe project, safe people, safe settings, safe data and safe output – to ensure data are protected.

"Cross state administrative data from the Coleridge training provided timely and relevant information for employers in the Midwest, allowing us to make a case for recruiting new firms to hire in the state."

- State Economic Development Official, November 2020

Over 500 people from more than 100 state agencies successfully completed an Applied Data Analytics program over the past three years. Together they initiated over 100 exploratory research projects focusing on critical policy concerns.

"Our data is the greatest untapped resource for improving public policy outcomes."

- State Workforce Director, August 2020

FOR MORE INFORMATION: **coleridgeinitiative.org/training/**