Democratizing Data: Sharing Data and Working Across Agency Boundaries

Nathan Barrett – Coleridge Initiative

COLERIDGE INITIATIVE

BIPARTISAN POLICY CENTER

Foundations for Evidence-Based Policymaking Act of 2018

The bipartisan Foundations for Evidence-Based Policymaking Act of 2018 builds off the work of the U.S. Commission on Evidence-Based Policymaking to strengthen data privacy protections, improve secure access to data, and enhance the federal government's capacity for producing and using evidence.

Strengthens Privacy Protections

Maintains Strong Confidentiality Protections for Sensitive Data. Reauthorizes the Confidential Information Protection and Statistical Efficiency Act (CIPSEA), an existing law that gives the American public strong privacy safeguards and legal protections for appropriate uses of confidential data.

Institutes Processes to Assess Data Risks. Strengthens efforts to protect confidentiality while making data accessible for evidence building and transparent to the public by requiring comprehensive risk assessments for certain publicly released data.

Enhances Public Trust in Data. Improves public trust in statistical activities by codifying language directing certain agencies to establish procedures to protect trust in data activities by appropriately maintaining objectivity, independence, and confidentiality.

Establishes Consistent Leadership on Key Data Issues. Ensures a senior leader in each agency is responsible for protecting privacy and ensuring confidentiality protections are appropriately applied by creating chief data officers.

Improves Secure Data Access

Encourages Agencies to Make Data Public and Open When Possible. Takes steps to improve the public information about what data government currently holds and make data publicly available when possible and in the public interest.

Requires Development of Data Inventories. Enables researchers and evaluators to better identify what government-collected data are available by directing agencies to create and maintain data inventories and publicly provide details about those datasets.

Makes Administrative Records Available for Evidence Building. Under a strong set of confidentiality protections, encourages that government data can and should be used to generate evidence about policies and programs, unless otherwise restricted by law.

Creates a Common Portal for Researcher Applications to Access Restricted Data. Reduces burden on researchers for applying to access government data by establishing a common application system for qualified individuals to access restricted, confidential data for approved projects.

Facilitates Continuous Feedback about Data Coordination. Promotes the use of data for evidence building by establishing a government advisory committee to review existing coordination and availability of data.

Enhances Government's Evidence Capacity

Directs Agencies to Develop Evidence Plans. Enables agencies to better prioritize evidence building by requiring that agencies document their key research questions, data needs, and planned activities.

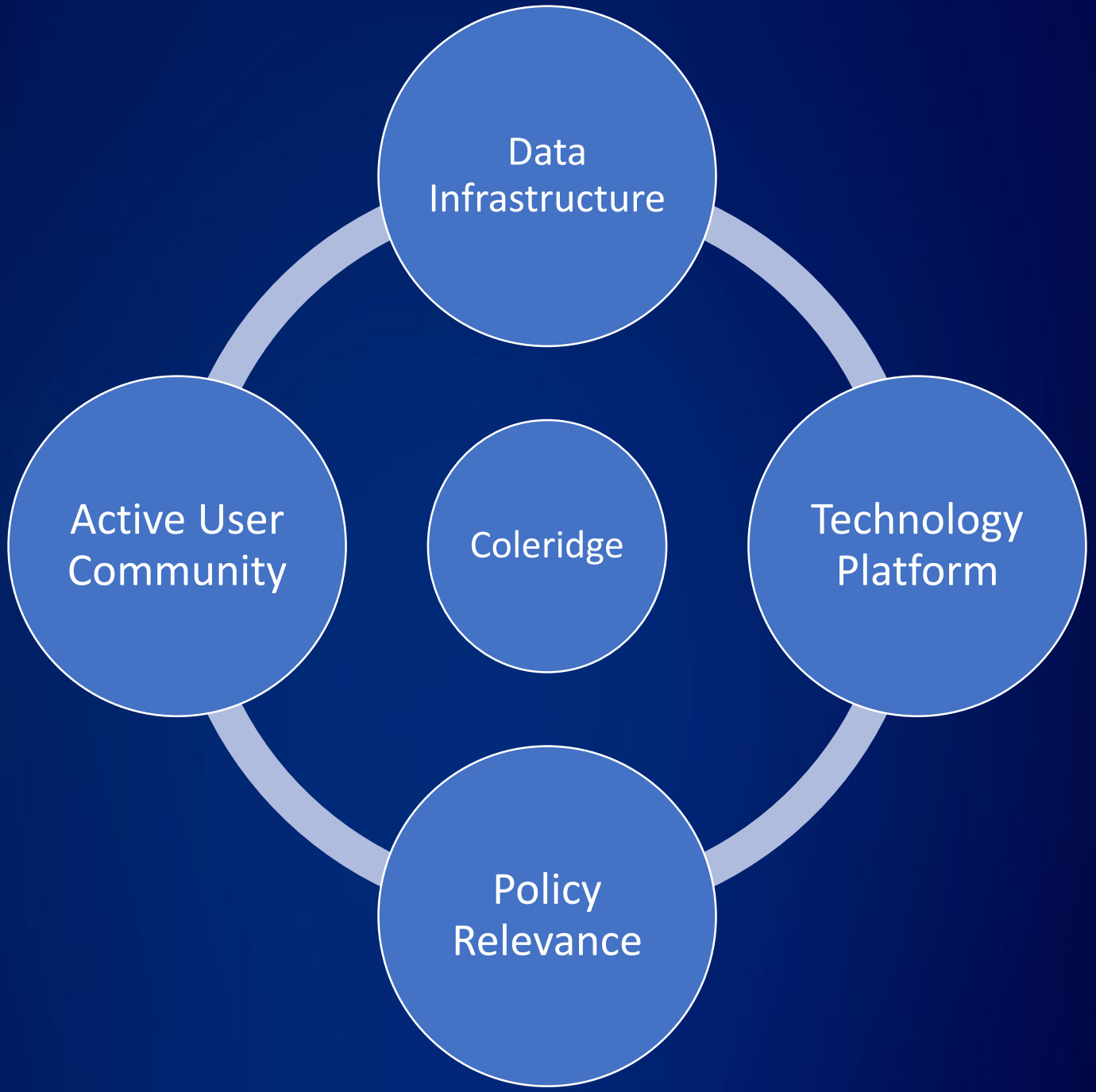
Prioritizes Evaluation Activities in Agencies. Improves agency capacity to engage in and use program evaluation by establishing evaluation officers in government agencies and requiring agencies to develop written evaluation policies.

Develops Baseline Information about the Resources Available for Evidence Building. Directs government agencies to periodically assess and report on their capabilities to engage in statistical, evaluation, and policy analysis activities and use the corresponding evidence for day-to-day government operations.

Learn more at bipartisanpolicy.org/evidence

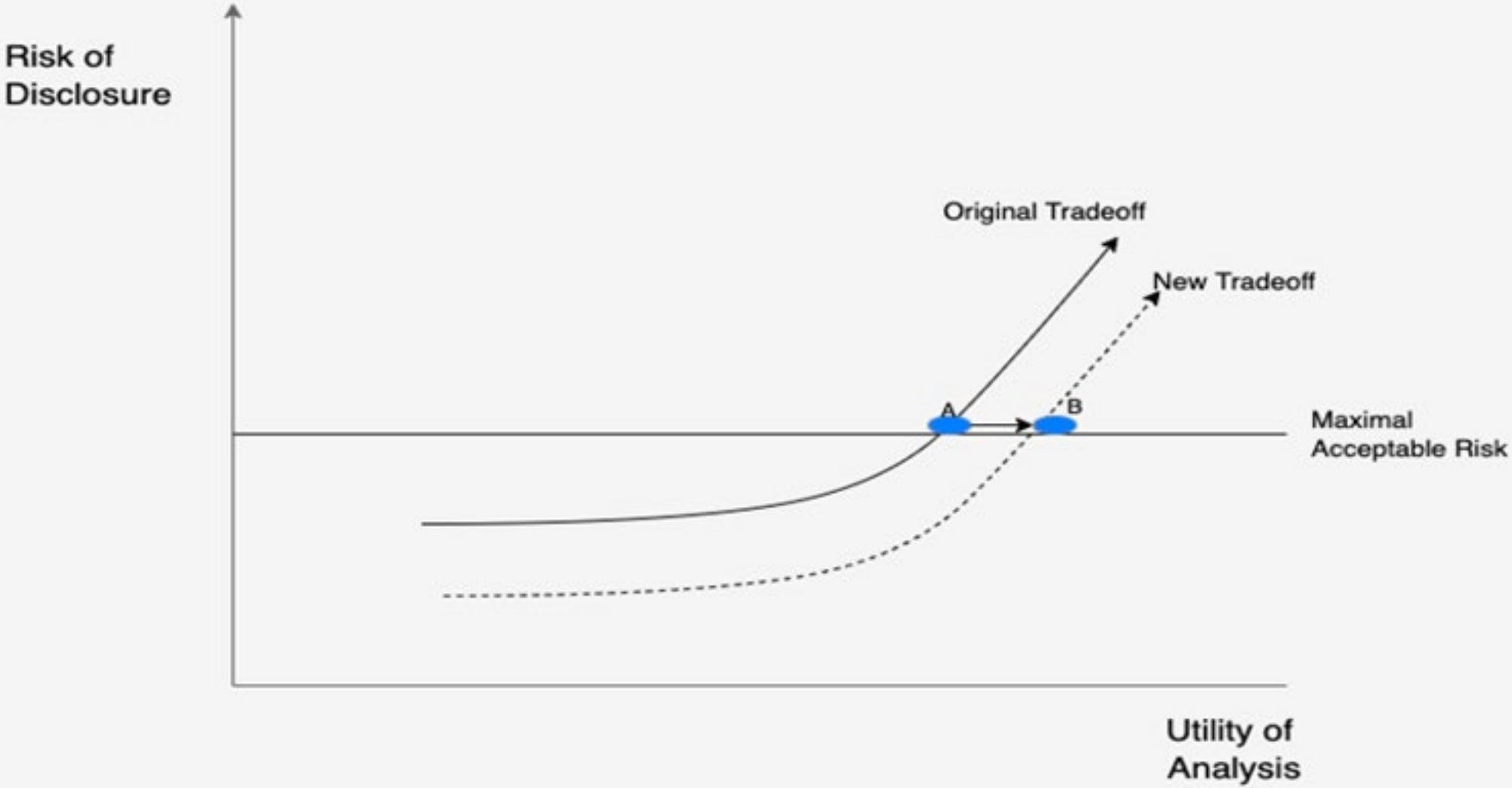
- Evidence-Based Policymaking Act
  - Strengthens Privacy Protections
  - Improves Secure Data Access
  - Enhances Government's Evidence Capacity
- States face common challenges
- The Coleridge Initiative: A non-profit organization, originally established at New York University, at the forefront of the '*Democratize Our Data Movement*'
- Mission: Enhance capacity of government agencies to develop data-based insights for policy
  - Technology
  - Collaboration
  - Training

COLERIDGE INITIATIVE



COLERIDGE INITIATIVE

Risk - Utility Tradeoff



Technology – Administrative Data Research Facility (ADRF)

ADRF

- Secure cloud-based data and computing environment that supports agencies and researchers in the development of evidence for policy and programs.
- FedRAMP Authorized
- Customizable configurations to meet various agency and researcher needs
- Operates under the five-safes framework

Five Safes Framework

Safe Projects

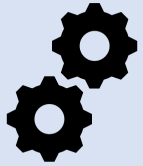
Appropriate Use of Data

- Only agency approved projects and data sets
- Only approved members can access the isolated project workspaces
- Controlled access to resources
- No shared environment between projects and resources

Safe People

Trained and Authorized Researchers

- Only approved researchers are permitted to access project workspaces
- User on-boarding process includes signing data use agreements, terms of use, security training module
- Data resources are explicitly granted based on project requirements
- Data resources are strictly in a read-only mode to ensure the integrity of the source data
- Security protocols follow strict FedRAMP guidelines



Safe Settings

Prevents Unauthorized Use

- Provides secure methods for agency micro-data transfer
- Only agency authorized personnel are invited to perform data transfers
- The transfer of data uses the FedRAMP Authorized, FIPS 140-2 validated, Kiteworks Secure Environment
- The transfer of data is restricted to upload operations only
- Additional security protocols include vulnerability scanning and third-party penetration testing

Safe Data

Protect Data Confidentiality

- Data Hashing - A custom stand-alone application simplifies and facilitates the hashing of data prior to transmission to the ADRF
- Data Stewardship - a web-based portal for data stewards to manage and monitor project and associated resources including project configurations, user activity, user onboarding status, and overall cost of a project on the ADRF etc.

Safe Exports

Non-Disclosive Exports

- Prevents users from unauthorized removal of any information within the secure environment
- Export requests are reviewed by data stewards following agency guidelines (e.g. proper cell suppression, no complementary disclosures, rounding and noise applied, no references to disclosive specific observations)
- Maintain a log of export requests for auditing purposes and to evaluate subsequent requests for complementary disclosure

The ADRF is a necessary but insufficient tool to support evidence-based decision-making

- How do we avoid The ADRF becoming a data mausoleum?
  - Trust
  - Active engagement
- **What questions can't you answer with the data and infrastructure you have available to you?**

Collaboration – Approaches and Outcomes

Regional Collaboratives

- The Southern, Eastern, and Midwest regional collaboratives help build the foundation of cross-state data sharing and products.
- Membership is comprised of agency leaders across policy domains, an administrative organization and Coleridge.
- Provide an opportunity for states to advise each other on issues of governance, promising products, grant opportunities, and pressing issues.
- Developing an RFI process to better engage the research community.

Democratizing Our Data Challenge

- The DDC was designed to develop and scale innovative product ideas by using government administrative data that are securely hosted in the Coleridge Initiative's Administrative Data Research Facility (ADRF).
- Vision is to transform the safe and secure use of data and evidence to inform policy in a fast-changing world.
- Has funded 10 projects with another round pending
  - Multi-state unemployment to reemployment
  - Multi-state postsecondary to workforce/Value Data Collaborative
  - K-12 to postsecondary to workforce

Legal and Technical Solutions

- Legal and technical solutions must support the linked data infrastructure we develop with state and federal partners.
- Legal:
  - Streamlined data sharing agreements that support amendments for approved work
  - Data stewards always maintain decision-making over access and approved projects
  - What do we do with screensharing?
- Technical
  - Shared folders across projects that allow for collaboration, learning, and efficiency
  - Documentation and code-sharing in Gitlab
  - Open ADRF
  - Redshift MPP
  - Closed environment creates some challenges

The ADRF and collaboration are necessary but insufficient to support evidence-based decision-making

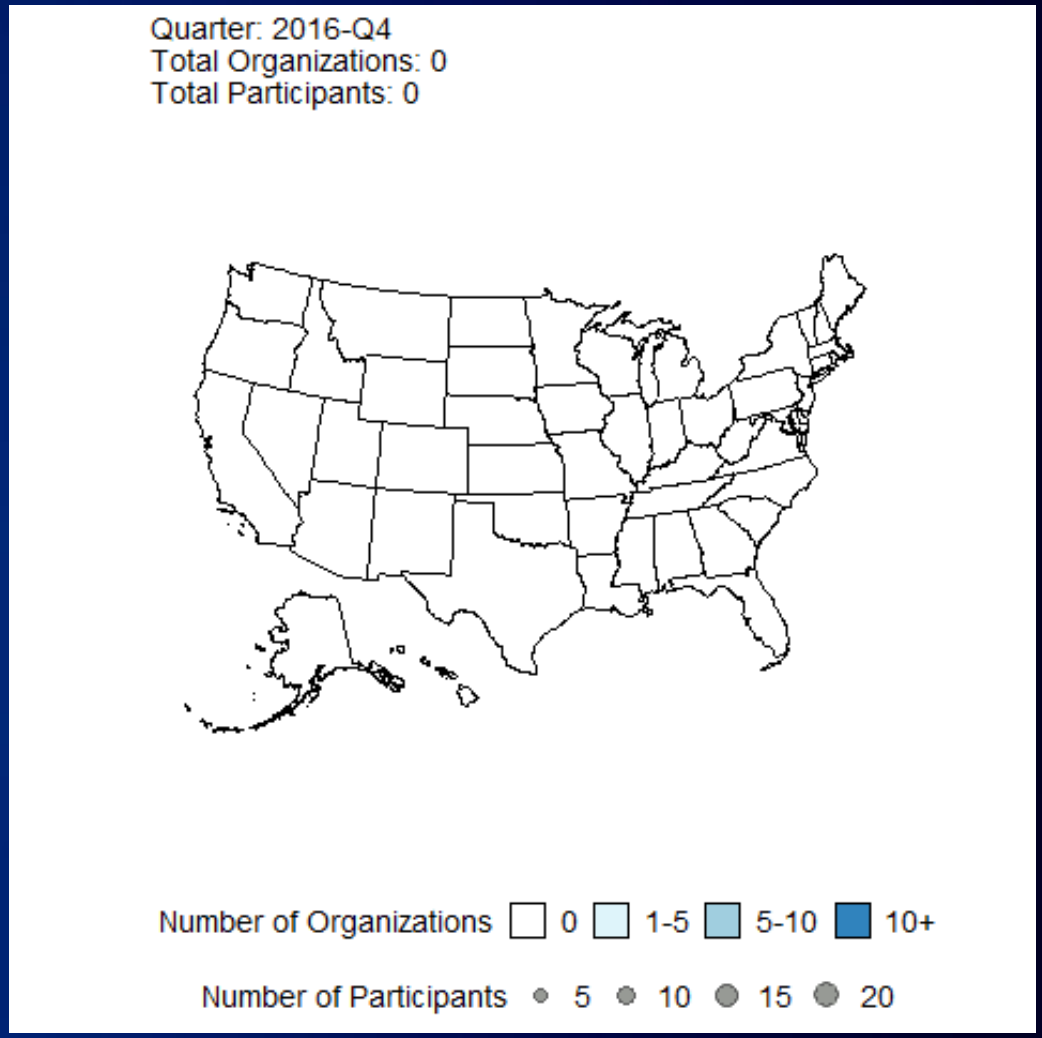
- How do truly democratize data?
- How do we institutionalize the work that the ADRF and the collaborative partnerships can support?
  - Integrated use
  - Product development
- **How do you know what questions you can ask and answer?**

Training

Our Work

- 31 trainings
  - 900+ participants
  - 300+ organizations
  - 40+ states

Training	Classes	Projects	Participants
Education to Workforce	16	79	391
Social Benefits and Workforce	5	34	158
Unemployment to Reemployment	4	24	114
Experiences of Formerly Incarcerated Individuals	2	28	125
Economic Development	2	17	72
Nutrition	1	6	21
Child Welfare	1	5	27
Total	31	193	908



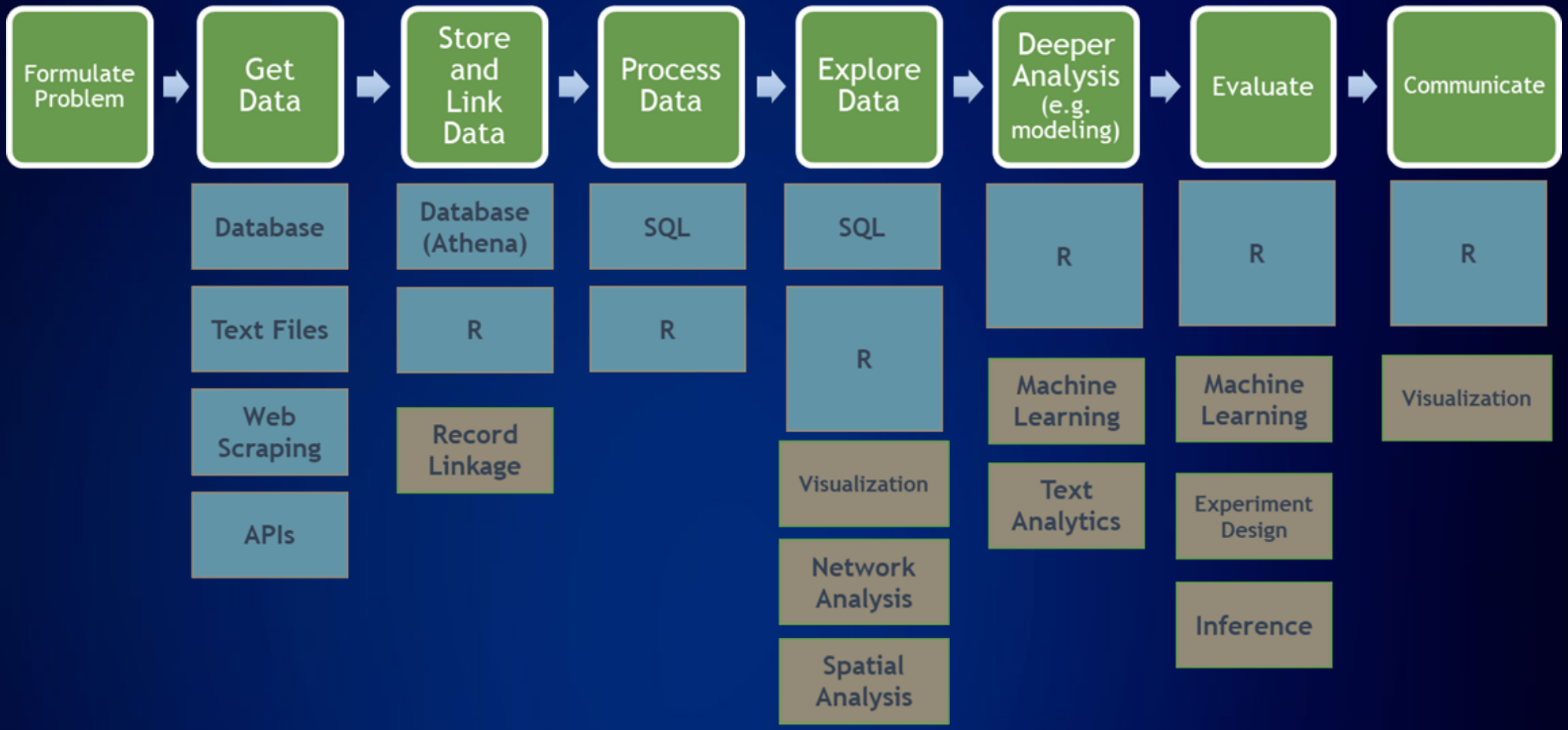
COLERIDGE INITIATIVE

Core Principles

- Every class is bespoke
- Topics best learned through a hands-on approach with actual micro-level administrative data
- Data science is a team sport
- Teams must represent multiple agencies and/or states
- Project-based - Teams develop their own research topic within the scope of the class
- Not a dissertation but the work must provide the foundation for something that is timely, relevant, and actionable
- The training can always be improved

Our Approach

- Development
  - Work with agency partner to establish the research question and required data
  - Project template
  - Build coding notebooks
  - Syllabus and lectures
- Delivery – Combination of lectures and facilitated team breakouts
  - Module 1 – Coding principles
  - Module 2 – Data preparation and exploration
  - Module 3 – Analytics
  - Final presentation


Formulate Problem
Get Data
Store and Link Data
Process Data
Explore Data
Deeper Analysis (e.g. modeling)
Evaluate
Communicate
Database
Text Files
Web Scraping
APIs
Database (Athena)
R
Record Linkage
SQL
R
SQL
R
Visualization
Network Analysis
Spatial Analysis
R
Machine Learning
Text Analytics
R
Machine Learning
Experiment Design
Inference
R
Visualization
Collaboration: "shared" folders in your project
Privacy, Confidentiality, Security

https://ada.coleridgeinitiative.org/
password: adrf

DAY 2

DATASET INTRODUCTIONS

- 55 mins: Dataset Introductions
- 5 mins: Break
- 55 mins: Group work
- 5 mins: Preview for Day 3

CLASS MATERIALS

Project Template

Slide Deck

Feedback Form

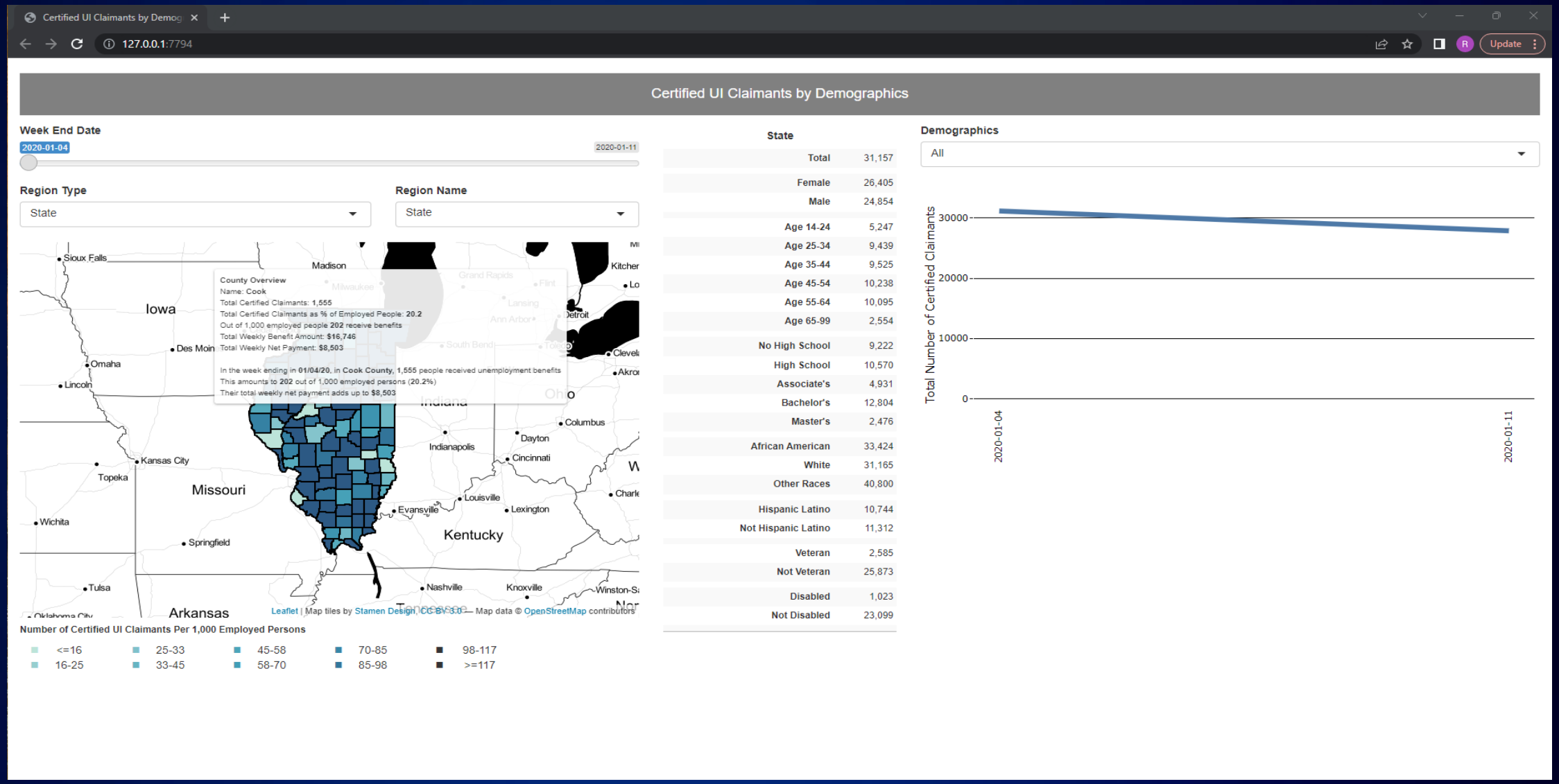
HOMEWORK FOR DAY 3

Exploratory Data Analysis Video / Slides

01_EDA.ipynb (ADRF)

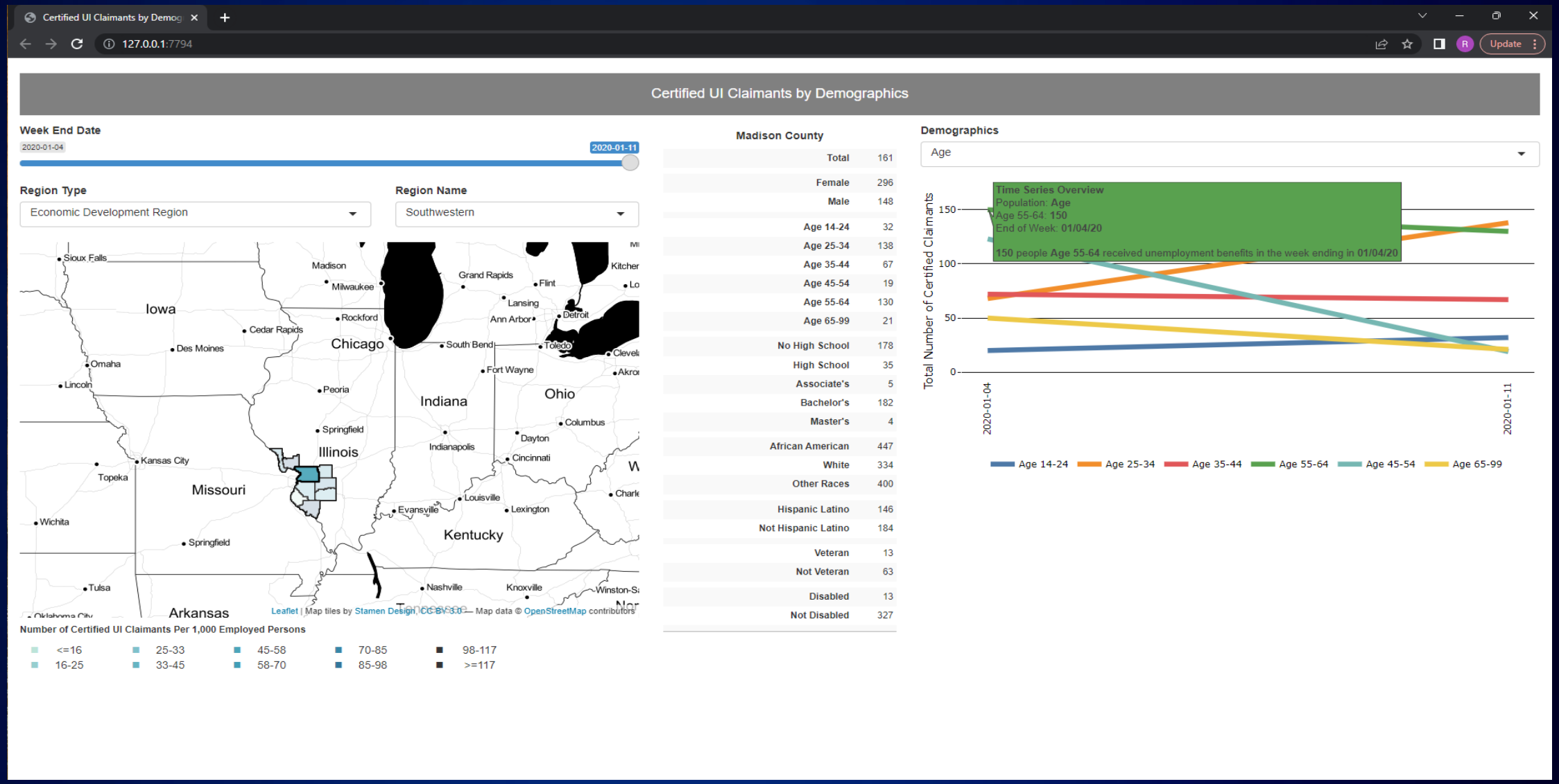
Discussion Question

Products from Trainings



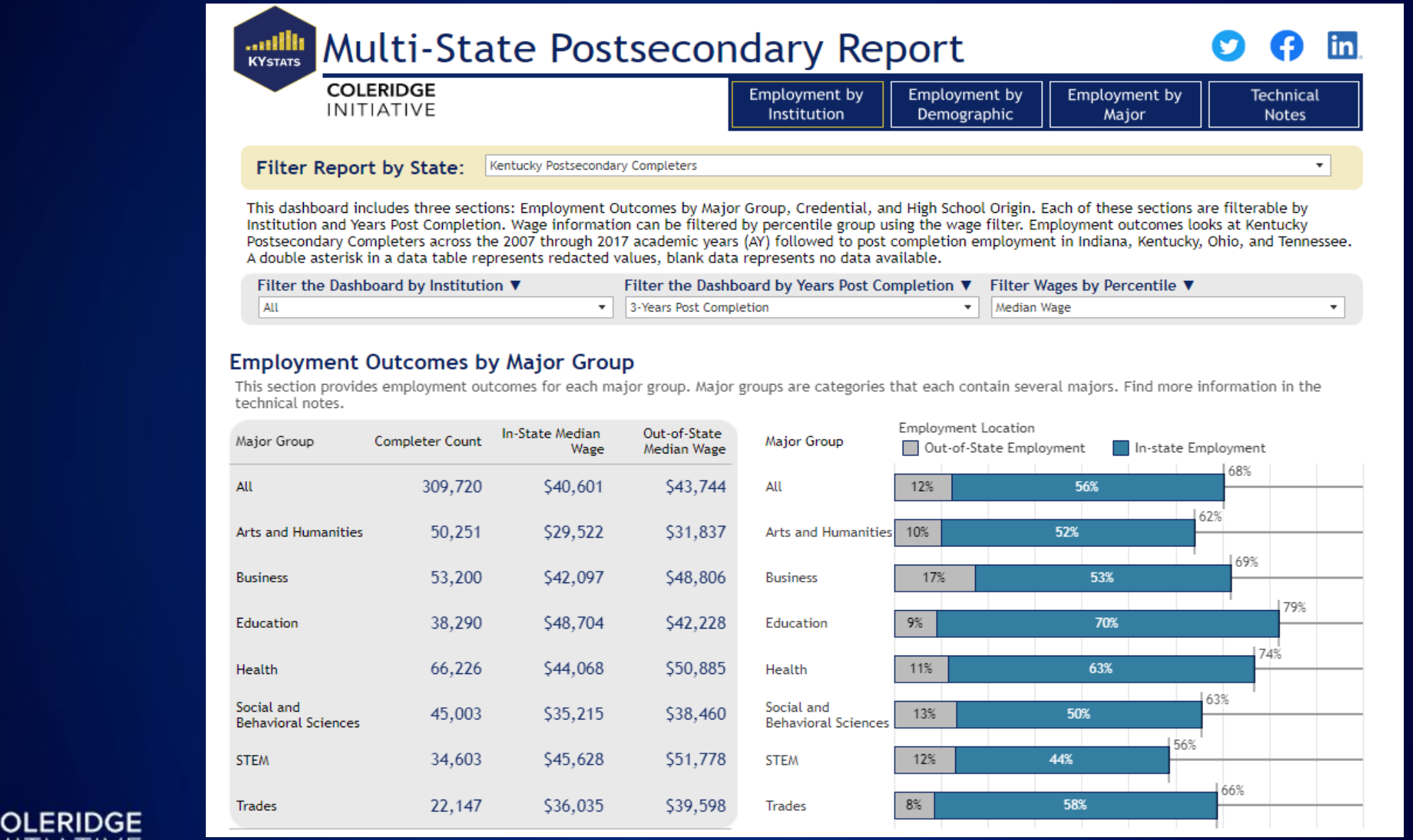
COLERIDGE INITIATIVE

Products from Trainings



COLERIDGE INITIATIVE

Products from Trainings



KYSTATS

Multi-State Postsecondary Report

COLERIDGE INITIATIVE

Employment by Institution | Employment by Demographic | Employment by Major | Technical Notes

Filter Report by State: Kentucky Postsecondary Completers

This dashboard includes three sections: Employment Outcomes by Major Group, Credential, and High School Origin. Each of these sections are filterable by Institution and Years Post Completion. Wage information can be filtered by percentile group using the wage filter. Employment outcomes looks at Kentucky Postsecondary Completers across the 2007 through 2017 academic years (AY) followed to post completion employment in Indiana, Kentucky, Ohio, and Tennessee. A double asterisk in a data table represents redacted values, blank data represents no data available.

Filter the Dashboard by Institution ▼: All

Filter the Dashboard by Years Post Completion ▼: 3-Years Post Completion

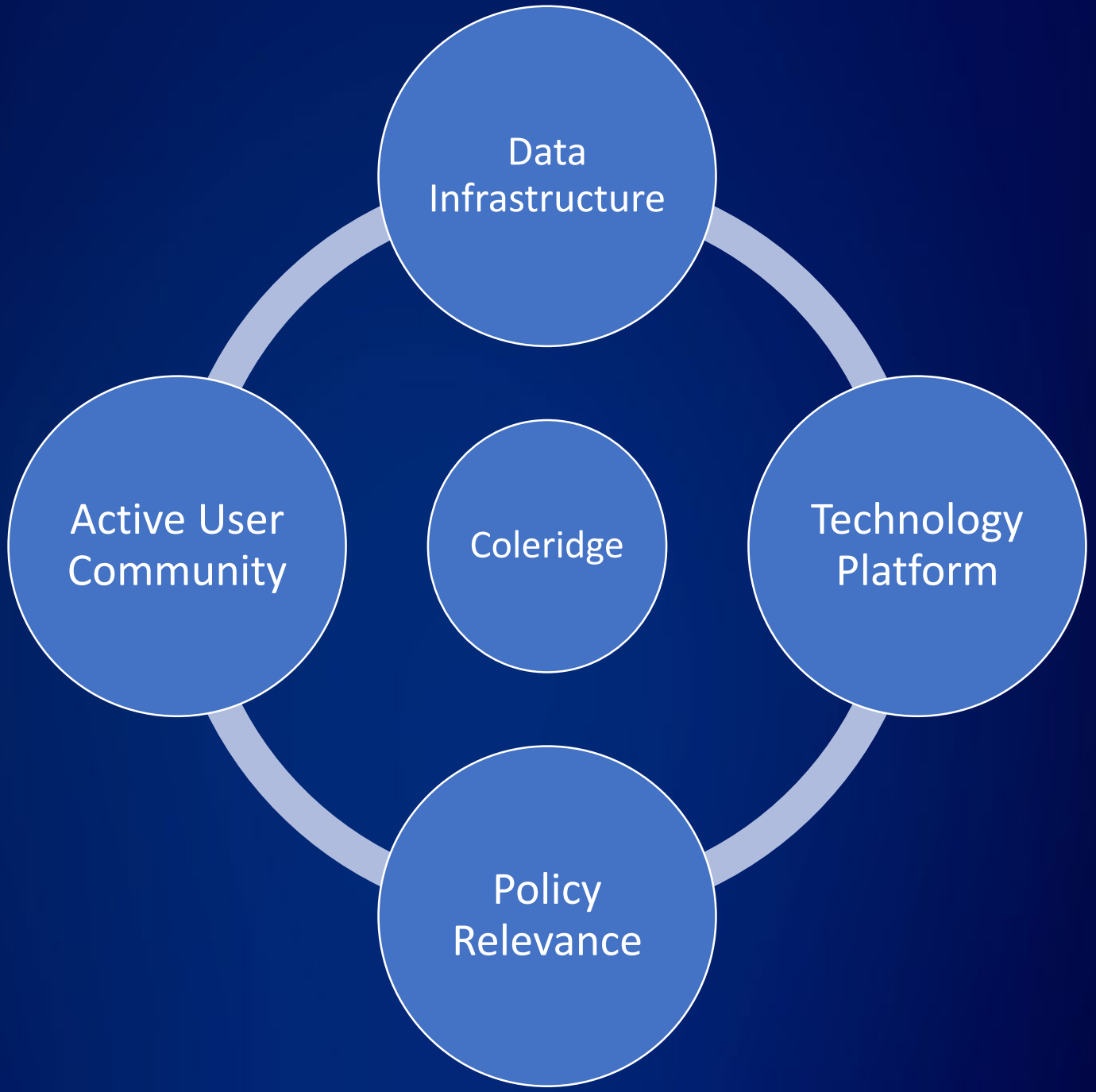
Filter Wages by Percentile ▼: Median Wage

Employment Outcomes by Major Group

This section provides employment outcomes for each major group. Major groups are categories that each contain several majors. Find more information in the technical notes.

Major Group	Completer Count	In-State Median Wage	Out-of-State Median Wage
All	309,720	$40,601	$43,744
Arts and Humanities	50,251	$29,522	$31,837
Business	53,200	$42,097	$48,806
Education	38,290	$48,704	$42,228
Health	66,226	$44,068	$50,885
Social and Behavioral Sciences	45,003	$35,215	$38,460
STEM	34,603	$45,628	$51,778
Trades	22,147	$36,035	$39,598

COLERIDGE INITIATIVE



COLERIDGE INITIATIVE

Concluding thoughts

- The current system is a result of decades of decisions based on information at the time. Systems have become institutionalized, and they will take time to change
- Trust is the first step in building the data infrastructure necessary to solve our most complicated policy issues
- We need champions that understand and can communicate the value proposition
- We must build an active and diverse community of users/consumers of the data infrastructure and the evidence it can provide
- Evidence must be relevant, timely, and actionable so that we can best serve the public interest