Survey Management Challenges

- Panel 1:

Top three challenges organizations are encountering in technology and survey computing

- Panel 2:

Challenges and achievements in using AI and data science approaches

Panel 2: Challenges and achievements in using AI and data science approaches

This panel will discuss the challenges and achievements that organizations have encountered is applying AI and data science approaches to their survey/data management projects. While there is significant publicity about the application of AI and data science techniques, the level of sophistication and experience varies.   Panelists will explore strategic approaches, best practices, and examples of where they have employed these techniques, and cover lessons learned in doing so.

Panel 2: Challenges and achievements in using AI and data science approaches

Moderator: Jane Shepherd – Westat

Panelists:

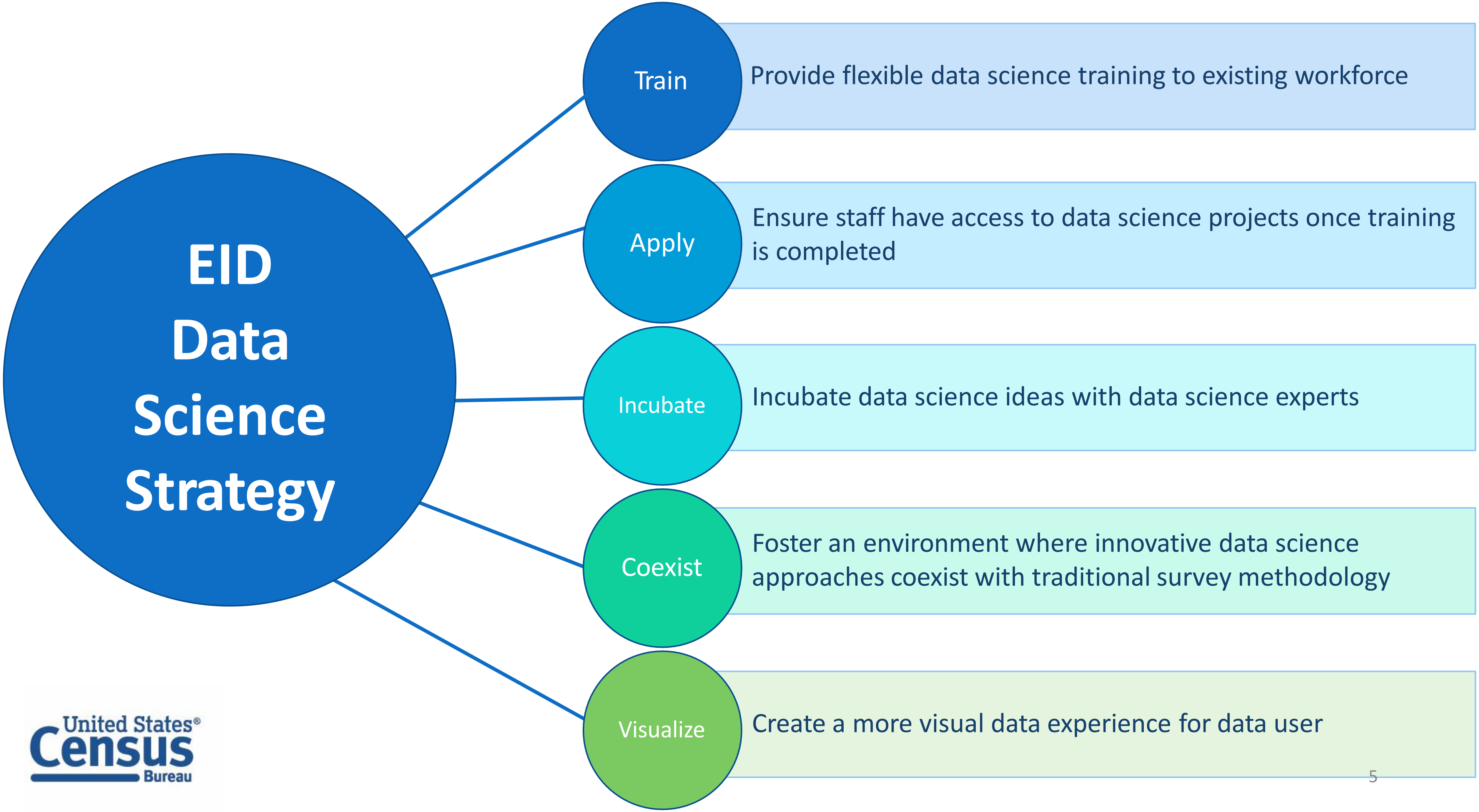
- Rebecca Hutchinson, U.S. Census Bureau
- Alex Measure, Bureau of Labor Statistics
- Jason Keller, NORC
- Gayle Bieler, RTI
- Marcelo Simas, Westat

The Census Bureau's Economic Indicator Division Data Science Strategy

Rebecca Hutchinson

April 13, 2021

Disclaimer: Any views expressed are those of the author and not necessarily those of the United States Census Bureau.



EID Data Science Strategy

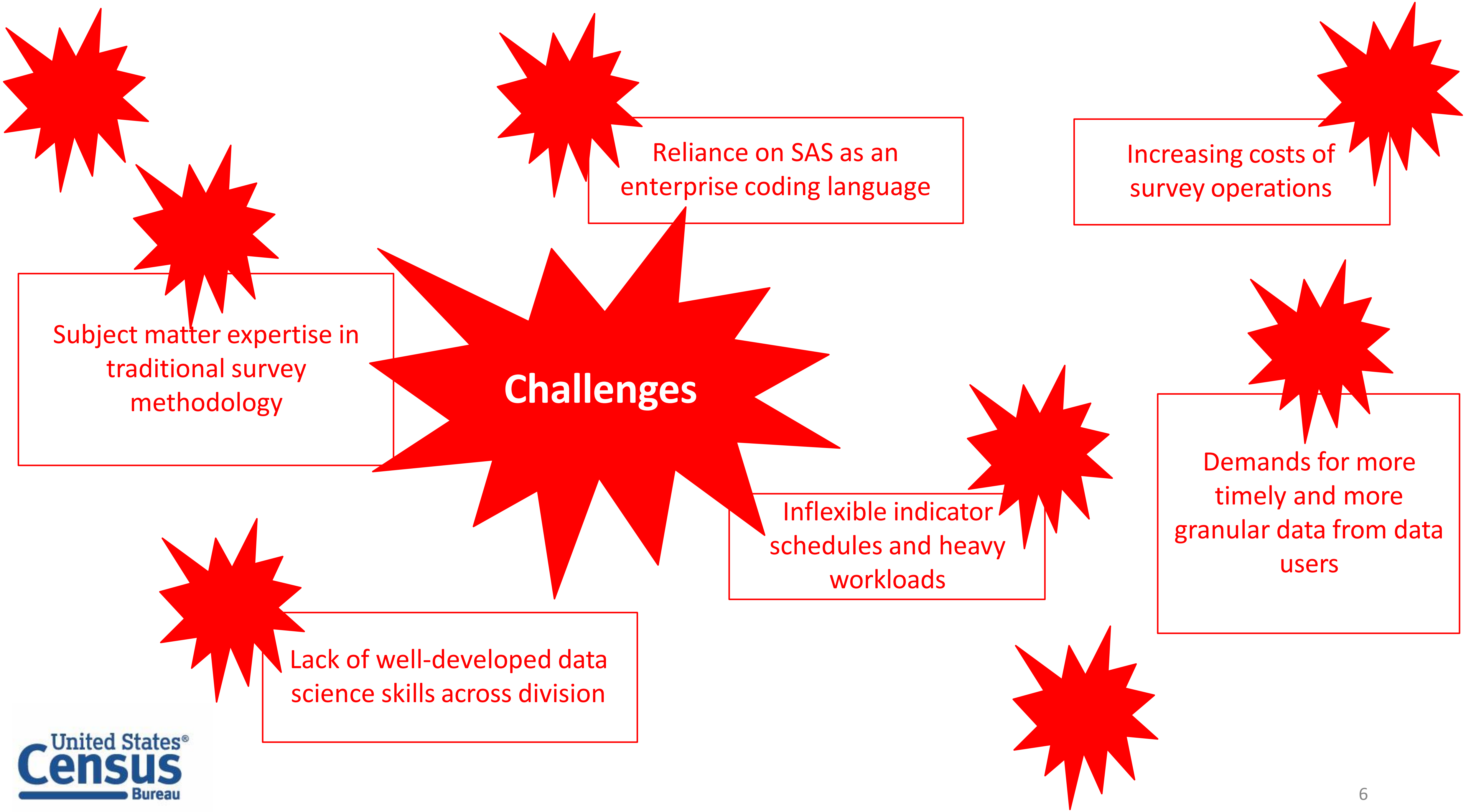
Train – Provide flexible data science training to existing workforce

Apply – Ensure staff have access to data science projects once training is completed

Incubate – Incubate data science ideas with data science experts

Coexist – Foster an environment where innovative data science approaches coexist with traditional survey methodology

Visualize – Create a more visual data experience for data user



Challenges

Subject matter expertise in traditional survey methodology

Reliance on SAS as an enterprise coding language

Increasing costs of survey operations

Demands for more timely and more granular data from data users

Inflexible indicator schedules and heavy workloads

Lack of well-developed data science skills across division

Data Science Training

Train

- Flexible training in Python, R, ArcGIS, Tableau, and SQL offered through online platforms Coursera and DataCamp.
- Allow staff up to five hours of work time each week to complete courses.
- Staff can either propose a project where they can utilize their new data science skills or they can be assigned to a new or existing project team.
- 30% of EID staff are enrolled in or have completed the training since 2018.
- Program has expanded to all of the Economic Directorate and over 200 staff have participated.

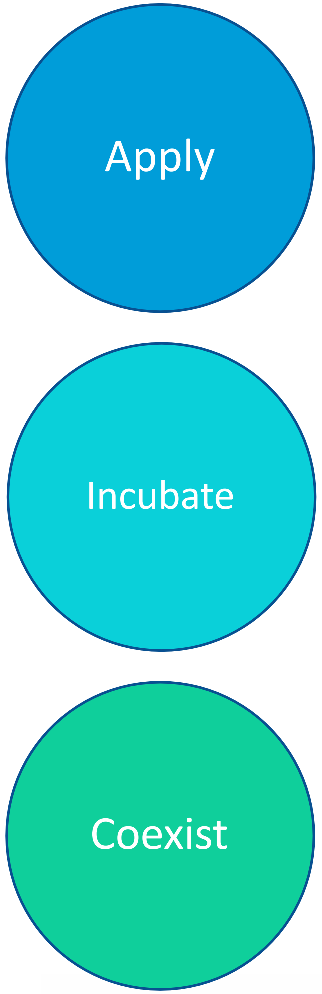
Construction Classification Work

Apply

Coexist

- Were only able to sample construction projects at a high level of construction type and more detailed codes are manually assigned only to the sampled construction projects.
- Implemented a classification machine learning algorithm to automate the assignment of construction project codes with less manual intervention.
- Ability to code additional projects will enable the utilization of a larger pool of projects to improve the quality and granularity of the Construction Spending estimates.

SEC Work



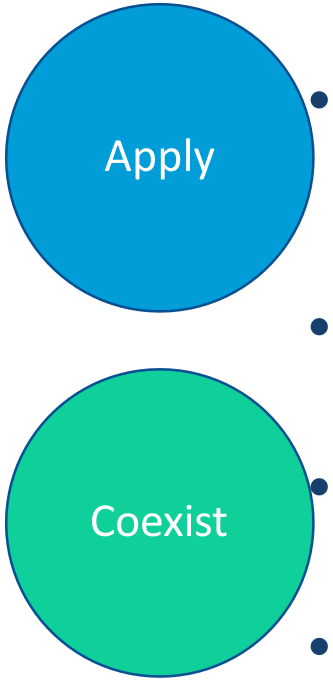
Apply

Incubate

Coexist

- Indicator staff rely heavily on publicly-available financial filings for the SEC. Extracting information from these forms is a manual effort. Piloted automating this process with a Civic Digital Fellow.
- Automated the scraping of SEC filings in XML filings.
- Developed process to automate the matching of SEC form items to survey form items.
- Implemented machine learning algorithm to predict and assign QFR items from SEC filings.

Improving NAICS Assignments



Apply

Coexist

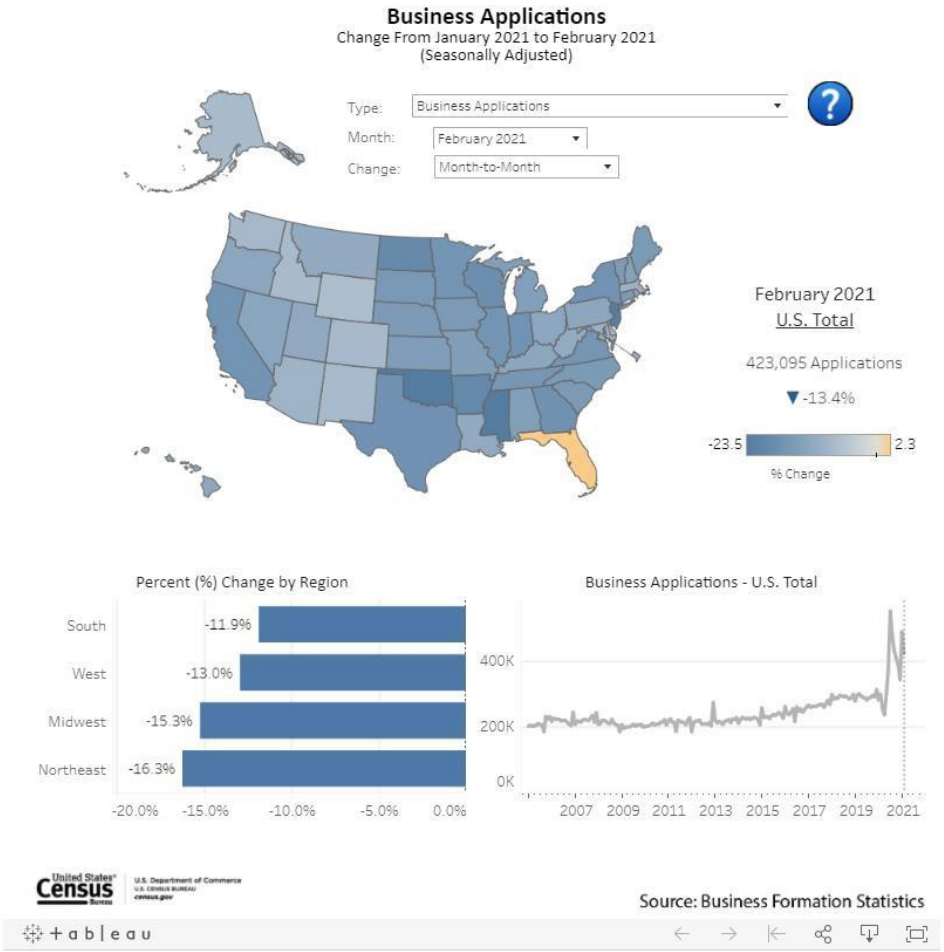
- Numerous North American Industry Classification System (NAICS) code machine learning efforts underway across the Census Bureau.
- Wanted to publish the Business Formation Statistics (BFS) by NAICS and needed to assess if the NAICS assignments done by the Census NAICS autocoder could be improved.
- Compared Census autocoder assignments for BFS records to assignments done by three NAICS ML technologies.
- Found that the Census autocoder does a good job; use one of the ML technologies to supplement the autocoder.
- First Monthly BFS NAICS release published in February 2021 utilizing this hybrid NAICS assignment approach.

Visualization

New state-level data products have allowed for a more visual data experience.

Apply

Visualize



Business Applications
Change From January 2021 to February 2021
(Seasonally Adjusted)

Type: Business Applications

Month: February 2021

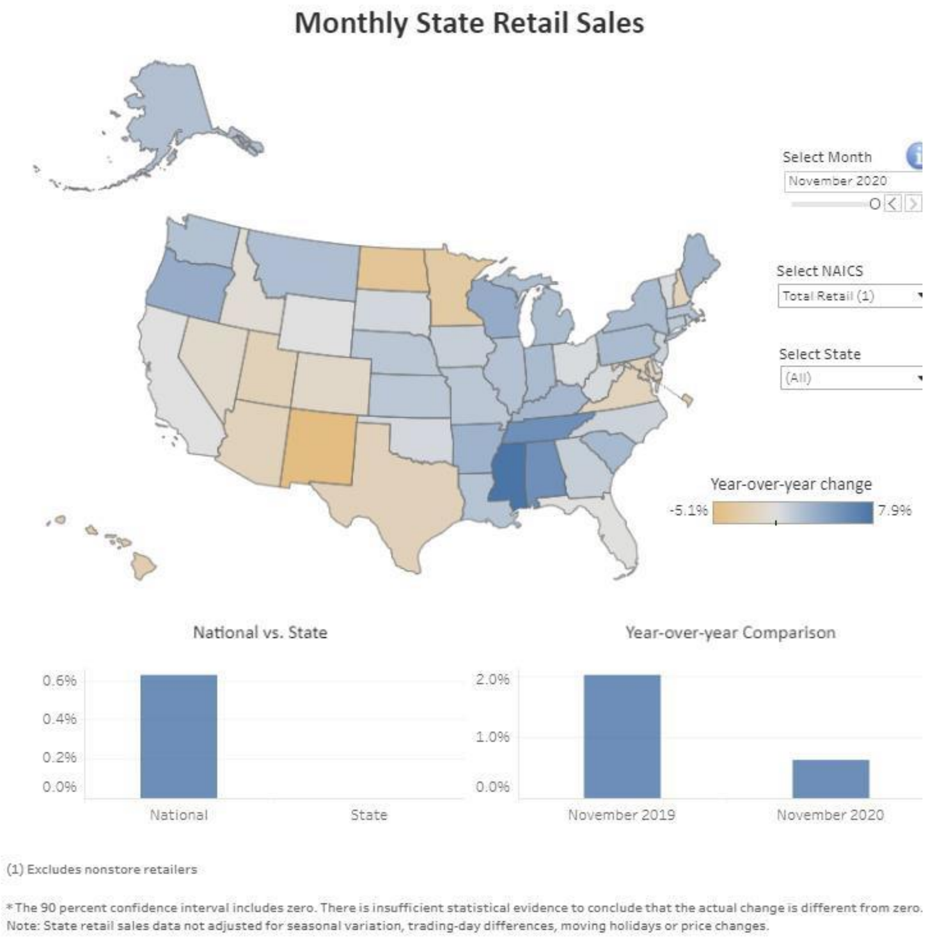
Change: Month-to-Month

February 2021
U.S. Total

423,095 Applications

▼ -13.4%

Source: Business Formation Statistics



Monthly State Retail Sales

Select Month: November 2020

Select NAICS: Total Retail (1)

Select State: (All)

(1) Excludes nonstore retailers

* The 90 percent confidence interval includes zero. There is insufficient statistical evidence to conclude that the actual change is different from zero.
Note: State retail sales data not adjusted for seasonal variation, trading-day differences, moving holidays or price changes.

Business Formation Statistics

Source: Business Formation Statistics - Data (census.gov)

Monthly State Retail Sales

Source: Monthly Retail Trade - State Retail Sales (census.gov)

Successes and Challenges with Machine Learning at BLS

Alex Measure

Success: Coding the Survey of Occupational Injuries and Illnesses

Example Narrative

Job title: sanitation worker

What was the employee doing just before the incident?
mopping floor in gym

What happened?
slipped on water on floor and fell

What part of the body was affected?
fractured right arm

What object directly harmed the employee?
wet floor

Codes Assigned

Occupation: 37-2011 (Janitor)

Nature: 111 (Fracture)

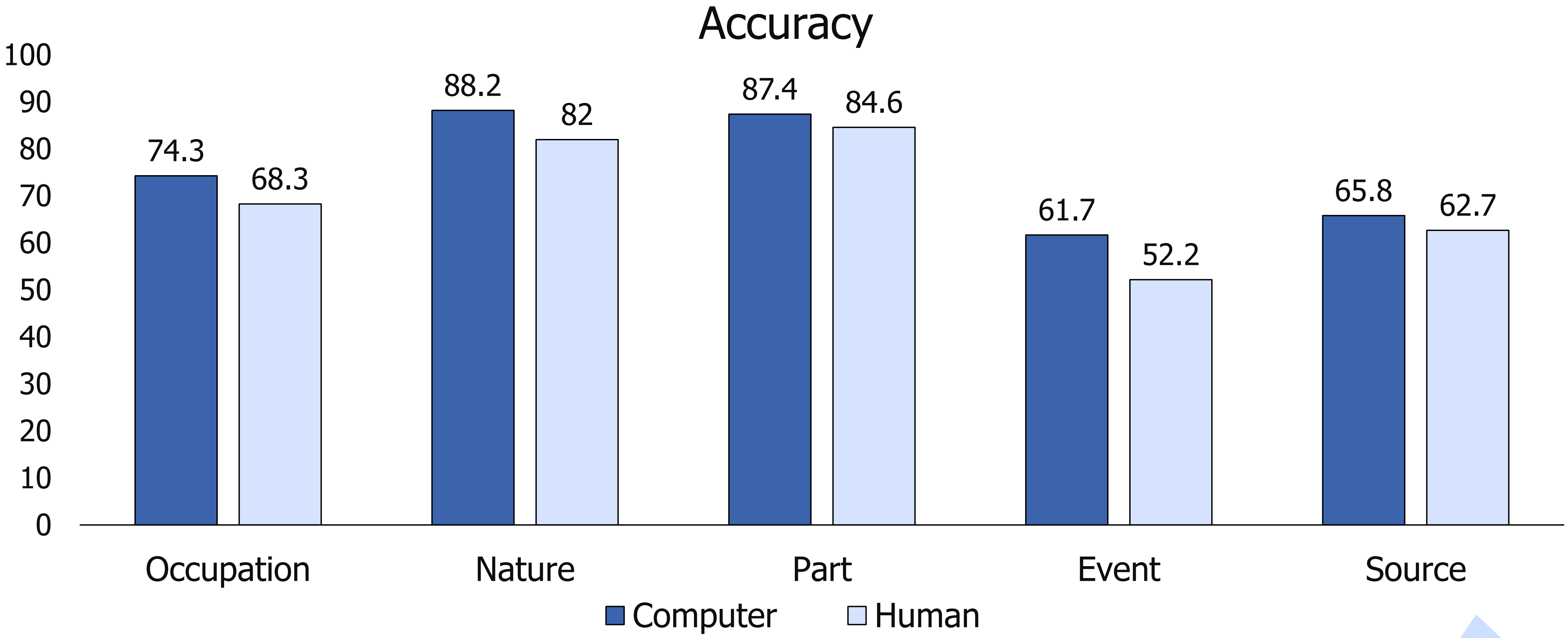
Part: 420 (Arm)

Event: 422 (Fall, slipping)

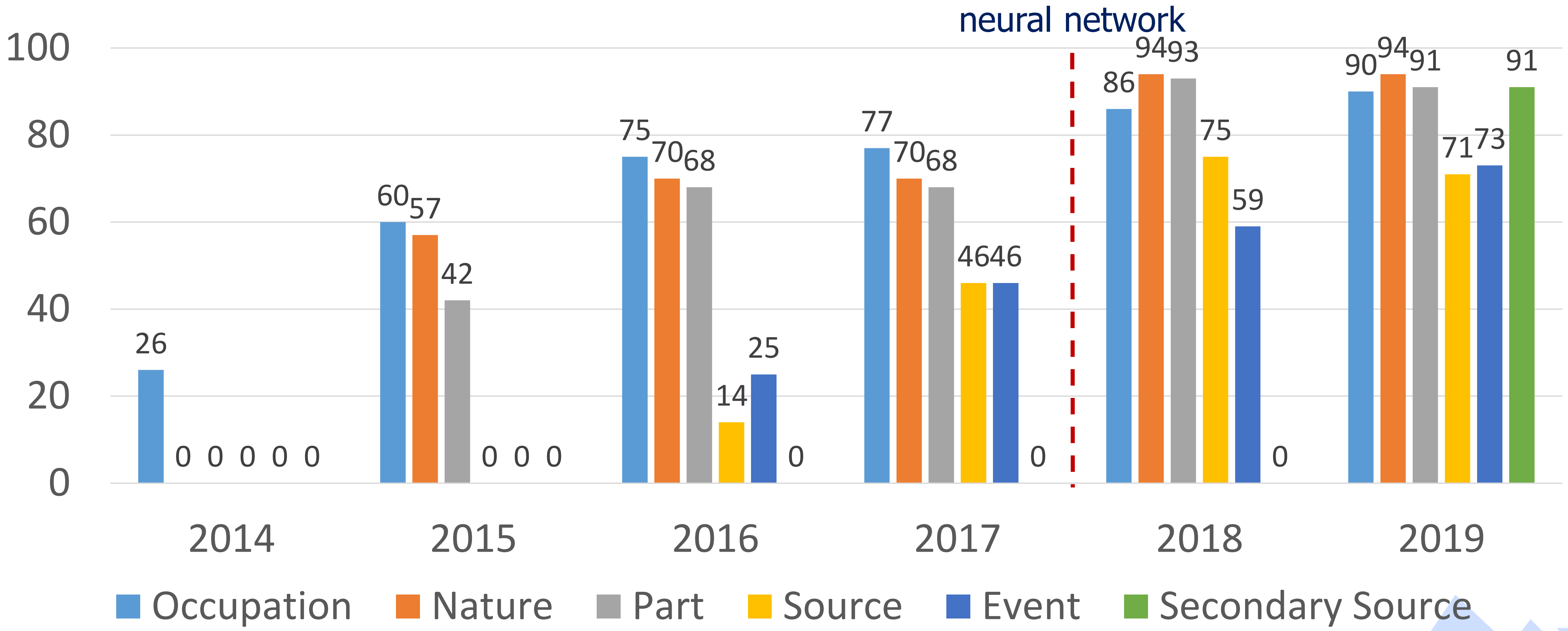
Source: 6620 (Floor)

Secondary: 9521(Water)

Does it work?

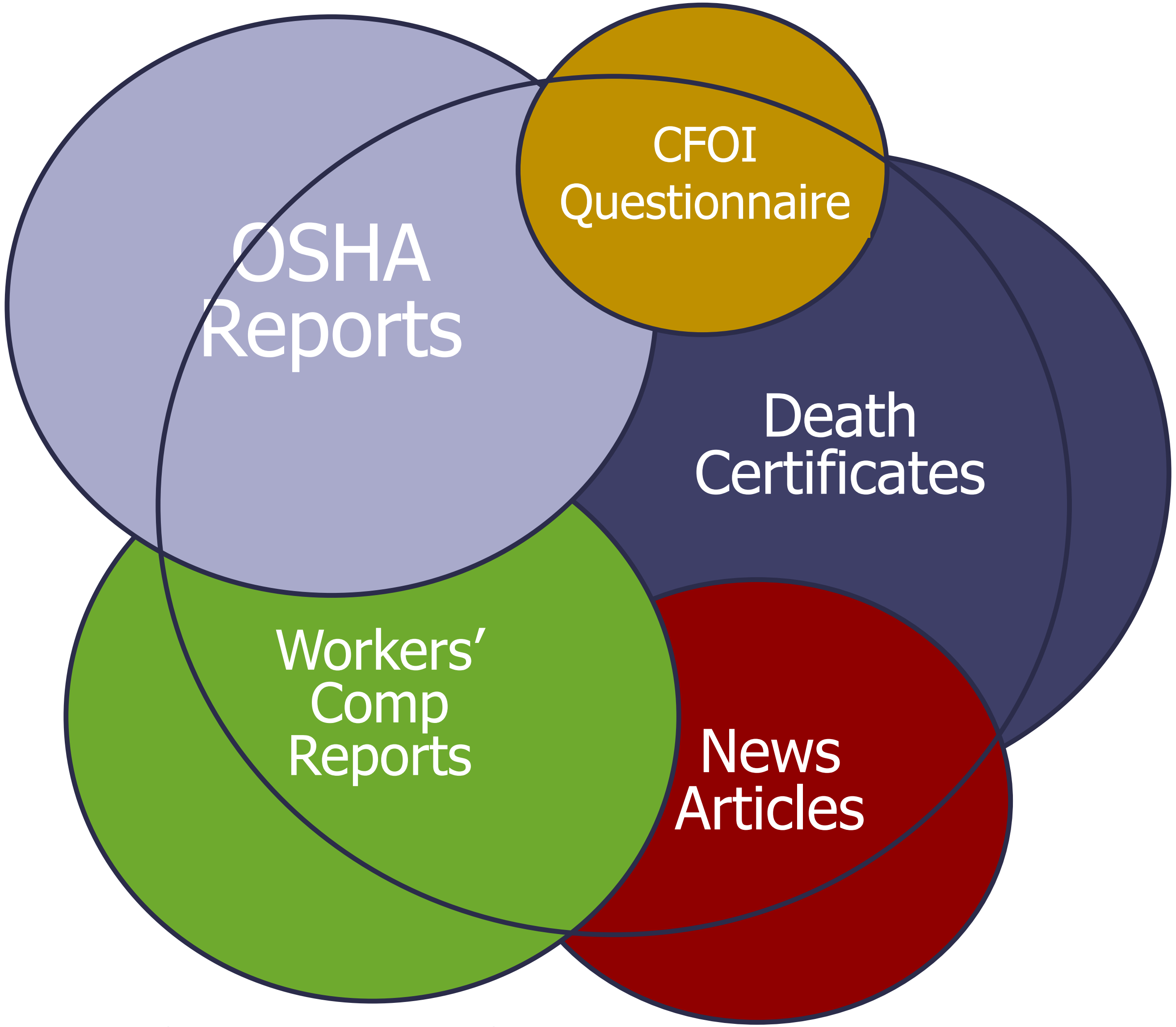

Accuracy
100
90
80
70
60
50
40
30
20
10
0
74.3
68.3
88.2
82
87.4
84.6
61.7
52.2
65.8
62.7
Occupation
Nature
Part
Event
Source
Computer
Human

SOII case coder (% autocoded)



Year	Occupation	Nature	Part	Source	Event	Secondary Source
2014	26	0	0	0	0	0
2015	60	57	42	0	0	0
2016	75	70	68	14	25	0
2017	77	70	68	46	46	0
2018	86	94	93	75	59	0
2019	90	94	91	71	73	91

Census of Fatal Occupational Injuries



- From diverse source docs
- Average ~5 per case
- Key challenges:
  - Sifting
  - Matching to others

Added Difficulty, Missing Names!

CFOI case file

Person	Company	Age	Narrative
XXXXXXX	XXXXXX	25	Car accident
Susan Carter	Tree Co.	74	Hit by tree
XXXXXXX	XXXXXX	34	Homicide

OSHA inspection file

Person	Company	Union	Industry
Suzy E. Carter	Joe’s Trees	Yes	124000
Frank Garcia	Cola Co.	No	332000
Jonathan Smith	A.C.M.E.	No	429000
Henry Long	BB Retail	Yes	620000

Success: Matching Records

- Even when no decedent name or company name present
  - Automatically identifies 92% of manual matches
  - Identifies hundreds of “missed” matches
  - Suggests “likely” matches for the rest
  - Enables automated consistency checking

- Now expanding to additional records
  - News articles
  - Death certificates

Where are all the production systems?

- Positive experience going back 9 years
- Surprisingly easy to build with modern tools
- Hundreds of similar tasks performed in BLS
- Where are all the ML systems?

Challenge #1: Skills

- Skills required
  - Machine learning algorithms
  - Natural language processing
  - Python programming

- Solution so far
  - Extensive re-skilling
  - Internal training programs

Challenge #2: Infrastructure

- Organization
- Hardware
  - Servers
  - GPU’s
- Tools
  - Version control
  - Issue tracking

- Solutions so far
  - Decentralized
  - New collaborations with IT
  - In-house GPU’s
  - BLS GitLab

Contact Information

Alex Measure
Senior Economist
Bureau of Labor Statistics
measure.alex@bls.gov

Challenges and Achievements in Using AI and Data Science Approaches

FedCASIC 2021

Jason Keller
Senior Data Scientist
NORC

NORC at the University of Chicago

Developing the Model

Using an off-the-shelf library is only part of the story. How do you evaluate the model? How do you know when it is good enough?

Cross Validation

- This is a default must.
- Think about the class and label distribution (stratification).
- Think about metrics, and look at more than one (F1, confusion matrix, etc.).

Grid Search for Parameters

- Running cross validation over many parameters.
- This should not be done willy-nilly.
  - **Think about what the parameters do, and which have the most impact.**
- Go from a broad to a narrow window

Model Evaluation and Development.

Dummy classifier

- How well does the model compare to chance?
- Different definitions of “chance”: most frequent value, random weighted selection.

Multi-class/multi-label

- No law against multiple models.
- Consider going from broad class to narrower classes.
- Consider turning it into a binary problem (if the number of classes is manageable).

Logistic Regression

- For binary problems, do NOT discount the logistic regression.
- Tried and true, easy to understand, and provides a good baseline for other models.

Using the Model in Production

Developing the model is only part of it.

The pitfalls of the traditional manual process

- It is all too easy to fall into the trap of the fully-manual process
  - **Someone gives you a flat file -> You run your magical code -> You provide the output.**
- Manual processes just ask for trouble in the long run
  - **Easy to make mistakes.**
  - **Single point of failure**
  - **What if you are not the one running it? What if the model is going to the client?**
  - **What if the mechanics of the survey depend on the model, or is part of some other process?**

Putting the model to work

Moving to a production workflow

- Serialization!
  - Save the model to a file format you can hand to someone else (i.e. a pickle file in Python).
  - **Challenge:** Requires some level of technical knowledge from the user.
  - **Challenge:** May require intimate knowledge on how the model was developed (supporting libraries/classes).
- Containers!
  - Hand over the model environment in a neat package (Docker).
  - **Challenge:** Requires some level of technical knowledge from the user (This again?!).
  - **Challenge:** Not every IT department is equipped to support this.
- REST APIs!
  - Stick the model behind a web endpoint.
  - If hosting for the client, can abstract away all the technical requirements.
  - **Challenge:** Now there is one more thing to support and maintain.
  - **Challenge:** Reinventing the wheel; No ready-made, standard tool to do this, outside the cloud.

Thank you.

Jason Keller
Senior Data Scientist
Keller-Jason@orc.org

Research You Can Trust™

NORC at the University of Chicago

Challenges and Achievements Using AI and Data Science

FedCASIC Workshops
April 13-14, 2021

Gayle S. Bieler
Senior Director
Center for Data Science
RTI International

Center for Data Science: We Deploy Data Science into RTIs Research Business…

Center for Data Science
7 years, 30 people
200+ Projects

A vibrant team with a compelling social mission…
Data Science for Social Good

Solving important national problems, improving our local communities, and transforming scientific research

Holistic Take on Data Science Practice

Integration is Key

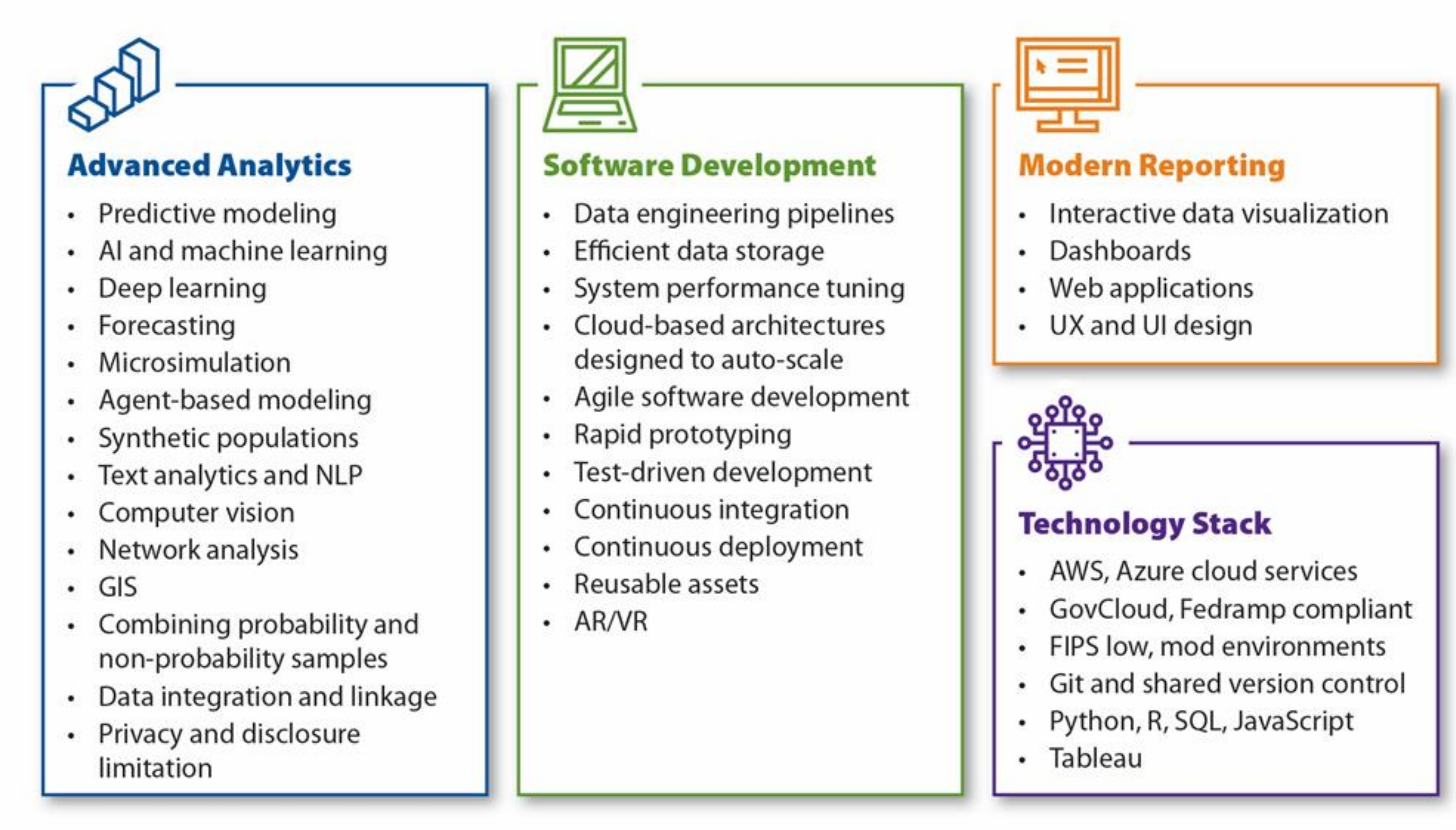
Data scientists…
Analytical work

Software developers
Data engineers…
Backbone of a data science team

Visual designers…
Visual communication of complex concepts

Backgrounds in other fields a huge plus

A Tip-to-Tail Practice



Advanced Analytics

- Predictive modeling
- AI and machine learning
- Deep learning
- Forecasting
- Microsimulation
- Agent-based modeling
- Synthetic populations
- Text analytics and NLP
- Computer vision
- Network analysis
- GIS
- Combining probability and non-probability samples
- Data integration and linkage
- Privacy and disclosure limitation

Software Development

- Data engineering pipelines
- Efficient data storage
- System performance tuning
- Cloud-based architectures designed to auto-scale
- Agile software development
- Rapid prototyping
- Test-driven development
- Continuous integration
- Continuous deployment
- Reusable assets
- AR/VR

Modern Reporting

- Interactive data visualization
- Dashboards
- Web applications
- UX and UI design

Technology Stack

- AWS, Azure cloud services
- GovCloud, Fedramp compliant
- FIPS low, mod environments
- Git and shared version control
- Python, R, SQL, JavaScript
- Tableau

We Solve Problems Across RTI: From the Social Sciences to the Lab Sciences

Research Domains

- Justice
- Public Health
- Clinical studies
- Environment
- Education
- Surveys
- Lab sciences (sensors)

Data Science Tools

- Open-source projects
- Public data products

For Whom?

NIJ

BJS

FDA

CMS
CENTERS FOR MEDICARE & MEDICAID SERVICES

NIH

CDC

NCDHHS
COVID-19 Response

DTRA
DEFENSE THREAT REDUCTION AGENCY
Detect. Deter. Defeat.

epa

DURHAM
1869
CITY OF MEDICINE

NCHS

DEPARTMENT OF COMMERCE
UNITED STATES OF AMERICA

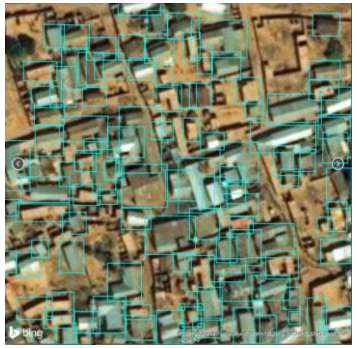
Greensboro Police

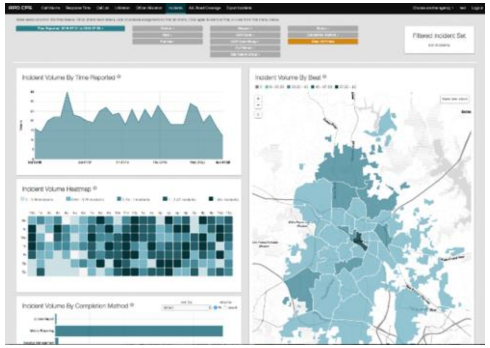
Arnold Ventures

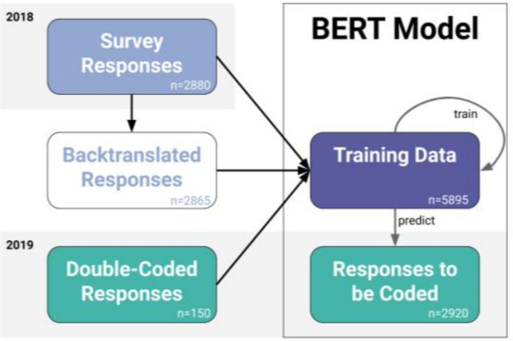
BILL & MELINDA GATES foundation

Survey Applications

- **Sample Design: Frames**
  - Computer vision using satellites and drones
- **Data Sharing and Dissemination**
  - Interactive dashboards
  - Interactive reports
  - Privacy
- **Data Processing with Text**
  - ***Verbatim Text Classification (NLP)***
    - Auto-coding (health, educ surveys)
  - ***Free-Text Survey Responses (NLP)***
    - Auto-coding
    - Informing survey question design



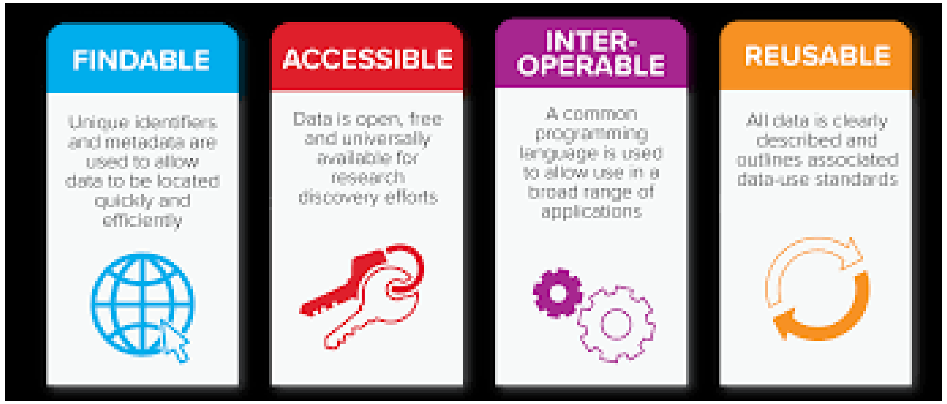




Data Management for Biological Applications (NIH)

Supporting FAIR Principles

- Cloud platforms
- Interactive data portals
- Data Sharing
- Data Dissemination
- Data Integration
- Data Harmonization
- Search Innovations



FINDABLE

Unique identifiers and metadata are used to allow data to be located quickly and efficiently

ACCESSIBLE

Data is open, free and universally available for research discovery efforts

INTER-OPERABLE

A common programming language is used to allow use in a broad range of applications

REUSABLE

All data is clearly described and outlines associated data-use standards

NIH National Heart, Lung, and Blood Institute | BioData CATALYST

NIH National Institute of Neurological Disorders and Stroke | mapMECFS

What Works: Team Values

Continuous learning and innovation

Data science needs to learn from itself

Practicing data science as a team sport

Team culture, Cross-disciplinary collaboration

Creating products and reusable assets that people use

Human-centered design, Agile software development

SMART

gobbli

What Works: Engaging with the World

Conferences
Client visits

Local Communities
Open-source Projects
Media Visibility
Publications
Blogs

gobbli

The Guardian

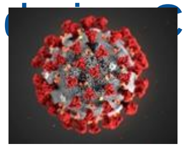
JAMA Pediatrics

nature

FiveThirtyEight

We Do Work That Matters

Informing Public Health Decisions in North Carolina: Rapid hospital and ICU capacity scenario modeling during COVID19



NCDHHS
COVID-19 Response

Helping federal agencies with Data Modernization:
Leveraging data as a strategic asset

RTI

February 25, 2020

Proposal to Enter into a Joint Venture Partnership with the National Technical Information Service for Data Innovation Support

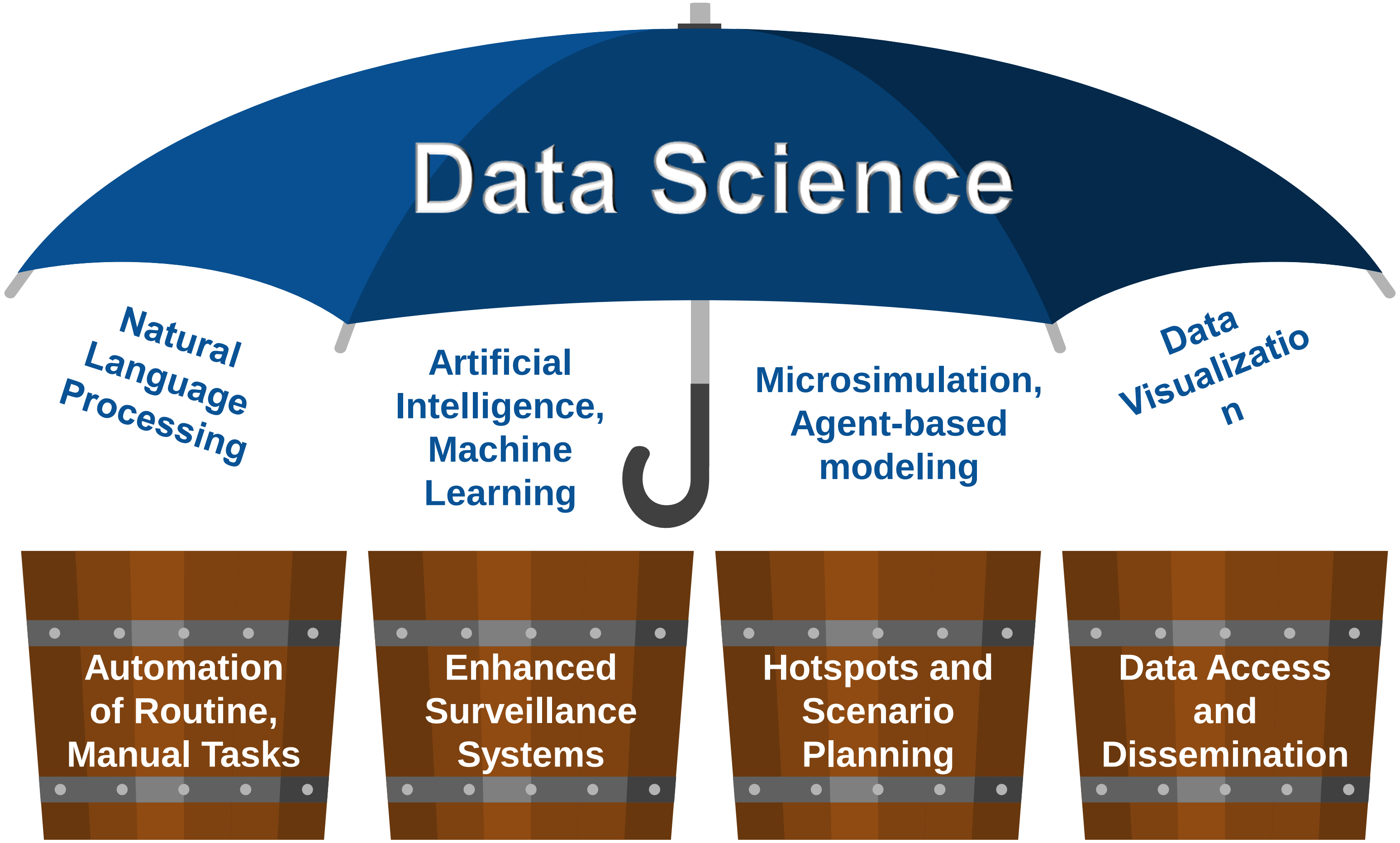
RTI Proposal No. 0282000.474
Federal Register Docket No. 191122–0088

Partnering with Crisis Text Line:
Firearm violence and suicide ideation during COVID19

CRISIS TEXT LINE |

CDC

Educating Our SMEs and Clients: The Data Science Umbrella

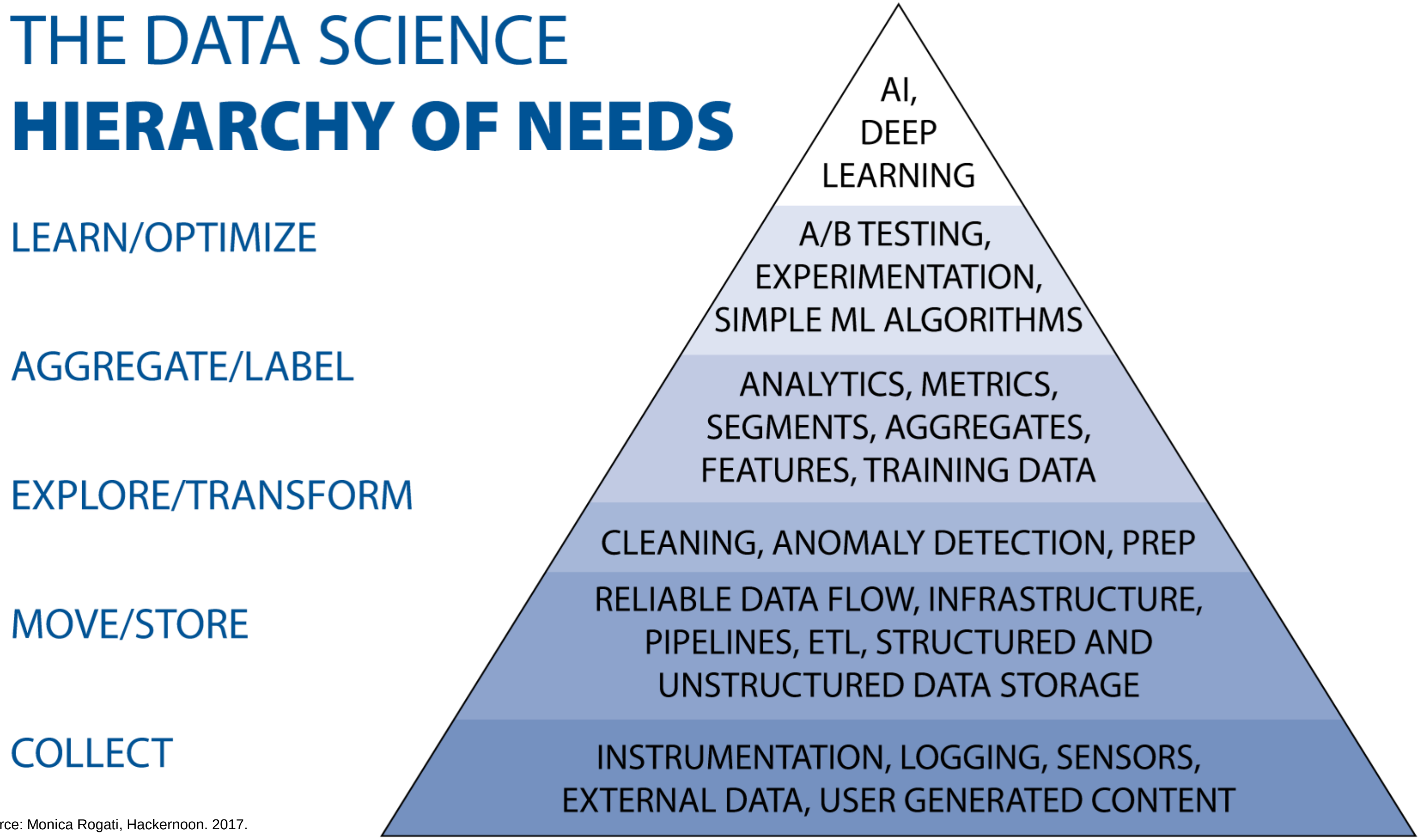

Data Science
Natural Language Processing
Artificial Intelligence, Machine Learning
Microsimulation, Agent-based modeling
Data Visualizatio n
Automation of Routine, Manual Tasks
Enhanced Surveillance Systems
Hotspots and Scenario Planning
Data Access and Dissemination

Connecting the Dots Through a Community of Practice

Data Science / AI Lunch and Learn Series

100-200 attendees per episode

Challenges: Level-Setting



THE DATA SCIENCE **HIERARCHY OF NEEDS**

- LEARN/OPTIMIZE
- AGGREGATE/LABEL
- EXPLORE/TRANSFORM
- MOVE/STORE
- COLLECT

AI, DEEP LEARNING

A/B TESTING, EXPERIMENTATION, SIMPLE ML ALGORITHMS

ANALYTICS, METRICS, SEGMENTS, AGGREGATES, FEATURES, TRAINING DATA

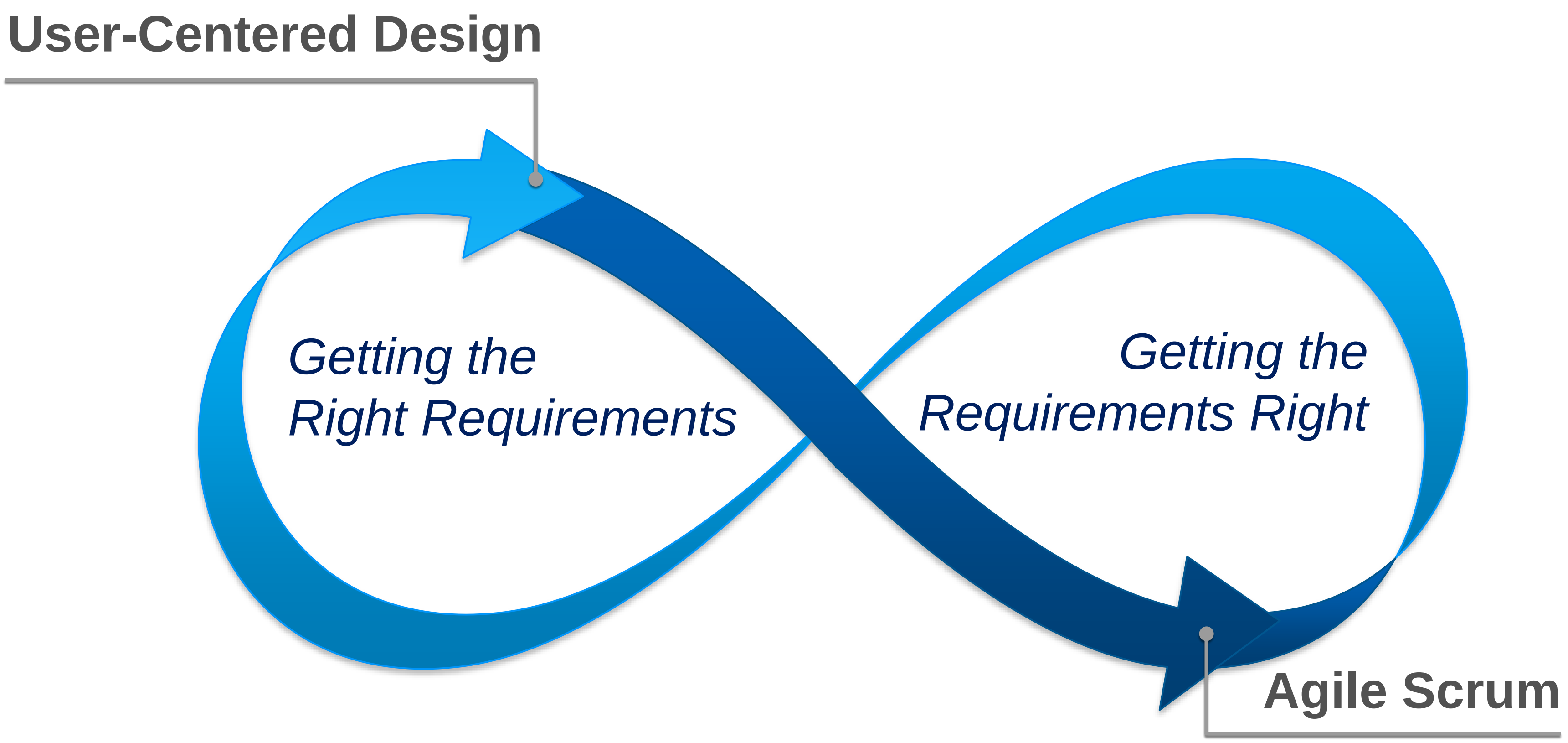
CLEANING, ANOMALY DETECTION, PREP

RELIABLE DATA FLOW, INFRASTRUCTURE, PIPELINES, ETL, STRUCTURED AND UNSTRUCTURED DATA STORAGE

INSTRUMENTATION, LOGGING, SENSORS, EXTERNAL DATA, USER GENERATED CONTENT

Source: Monica Rogati, Hackernoon. 2017.

Challenges: Embracing UCD and Agile Development



User-Centered Design

Getting the Right Requirements

Getting the Requirements Right

Agile Scrum

Thank You!

Applications of AI/ML in Survey Research

FedCASIC 2021

Marcelo Simas, Jiating (Kristin) Chen, Alex Cates

Artificial Intelligence and Machine Learning (AI / ML)

- Classification vs. Continuous Prediction Problems
  - What is this or what can you find in this?
    - Is this a hot-dog or a cat?
  - How much of this will I get? (like in a linear regression)
    - How hot will it be tomorrow?
- Learning Modes
  - Supervised – organize data ahead of time into groups and let the computer pick features and weight them
  - Unsupervised - let the machine (aka computer) identify groups (clustering)

Artificial Intelligence and Machine Learning (AI / ML)

- Artificial Intelligence (AI)
  - Nebulous and broad in scope
  - “Artificial intelligence is the science and engineering of making computers behave in ways that, until recently, we thought required human intelligence.” Andrew Moore
    - Example: computers can find and classify objects in images
  - More pragmatic -> if you’re using neural networks you are doing AI.

- Machine Learning (ML)
  - Algorithms that build models by following an automated process
  - More pragmatic -> if you’re not using neural networks and you can do what’s in the previous item you’re doing ML

But Really…

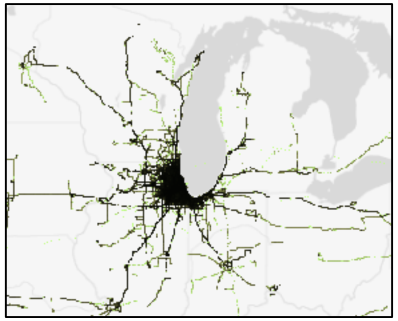
- “Stop Calling Everything AI”

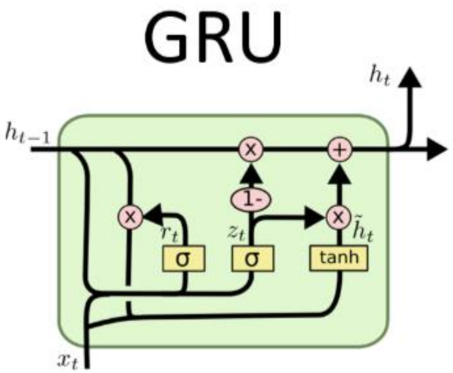
- “Artificial-intelligence systems are nowhere near advanced enough to replace humans in many tasks involving reasoning, real-world knowledge, and social interaction. They are showing human-level competence in low-level pattern recognition skills, but at the cognitive level they are merely imitating human intelligence, not engaging deeply and creatively”

  - Michael I. Jordan, University of California, Berkely, IEEE Spectrum (2021)
    - (source: https://spectrum.ieee.org/the-institute/ieee-member-news/stop-calling-everything-ai-machinelearning-pioneer-says)

AI Applications

- Imputation / Prediction of travel activity attributes in Chicago
- Westat’s DailyTravel smartphone app travel capture data
  - Confirmed places were used to train models for unconfirmed places
  - Used neural networks to predict travel mode from aggregated GPS and accelerometer data
  - User neural networks to predict trip purpose for unconfirmed places
  - Both networks used Gated Recurrent Units (GRUs)

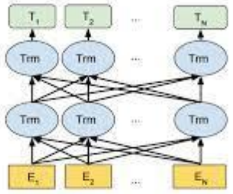




AI Applications

AHRQ Agency for Healthcare Research and Quality

- MEPS - classify open-ended field comments
  - Extracted keyword features using regex, named entities recognition and fuzzy string matching
  - Fine-tuned BERT to classify text into 10 categories
  - Deployed using on-prem using Python on k8s cluster

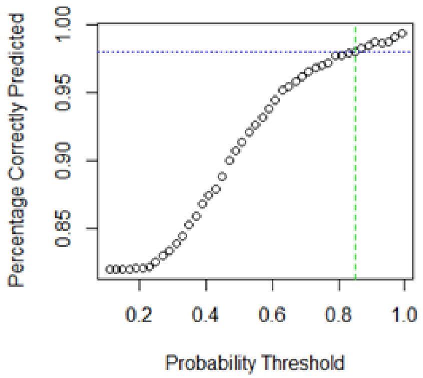


- DAWN – classify open-ended ED records to 8 drug abuse categories
  - Prepared text using GloVe word embeddings
  - Trained a multilayer bidirectional GRU with attention mechanism
  - Implemented in Python and deployed as cloud microservices

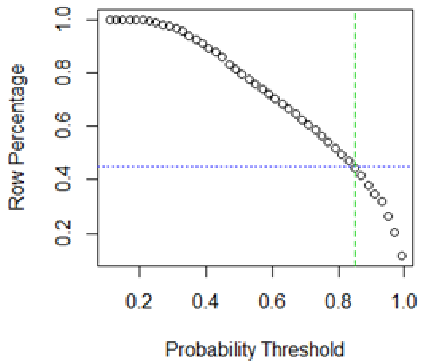
SAMHDA Substance Abuse & Mental Health Data Archive

Machine Learning Applications

- Suggestions within survey instruments for NAICS/SOC codes
  - Inversed Cosine Similarity Index between respondent text and category descriptions
  - R package is consumed by web surveys through OpenCPU



- Accelerate open end coding using NLP models in NHTS
  - Used manually up-coded open-end data on a subset of cases to train model that was then used in a shiny app to provide suggestions to coders
    - Selected a predicted probability threshold based on accuracy assessment

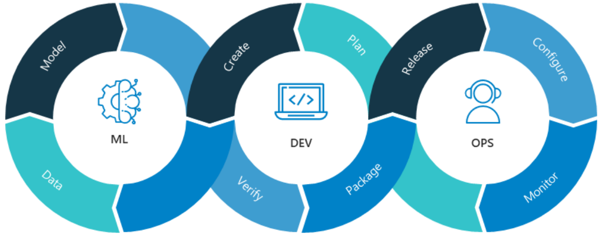


Machine Learning Applications

- Predict response propensity model for MEPS
  - Used paradata and demographic characteristics in sample frame
  - Trained a XGboost model
  - Created a R Package with model and supporting functions
- Also, indirectly by using procedures in packages such as **twang**

Challenges and Lessons Learned

- Most survey data is too thin for deep AI models which saw a lot of development when dealing with rich data
  - Free form text and multi (more than a few words)
- Data Science models need substantial IT support to integrate with production systems and data pipelines
  - Automated continuous deployment (MLOps)
  - Data privacy and PII require care when deploying models to the cloud and using open tools



- Rapidly evolving SOTA models introduce learning, application, and diagnoses challenges for data science practitioners

Discussion