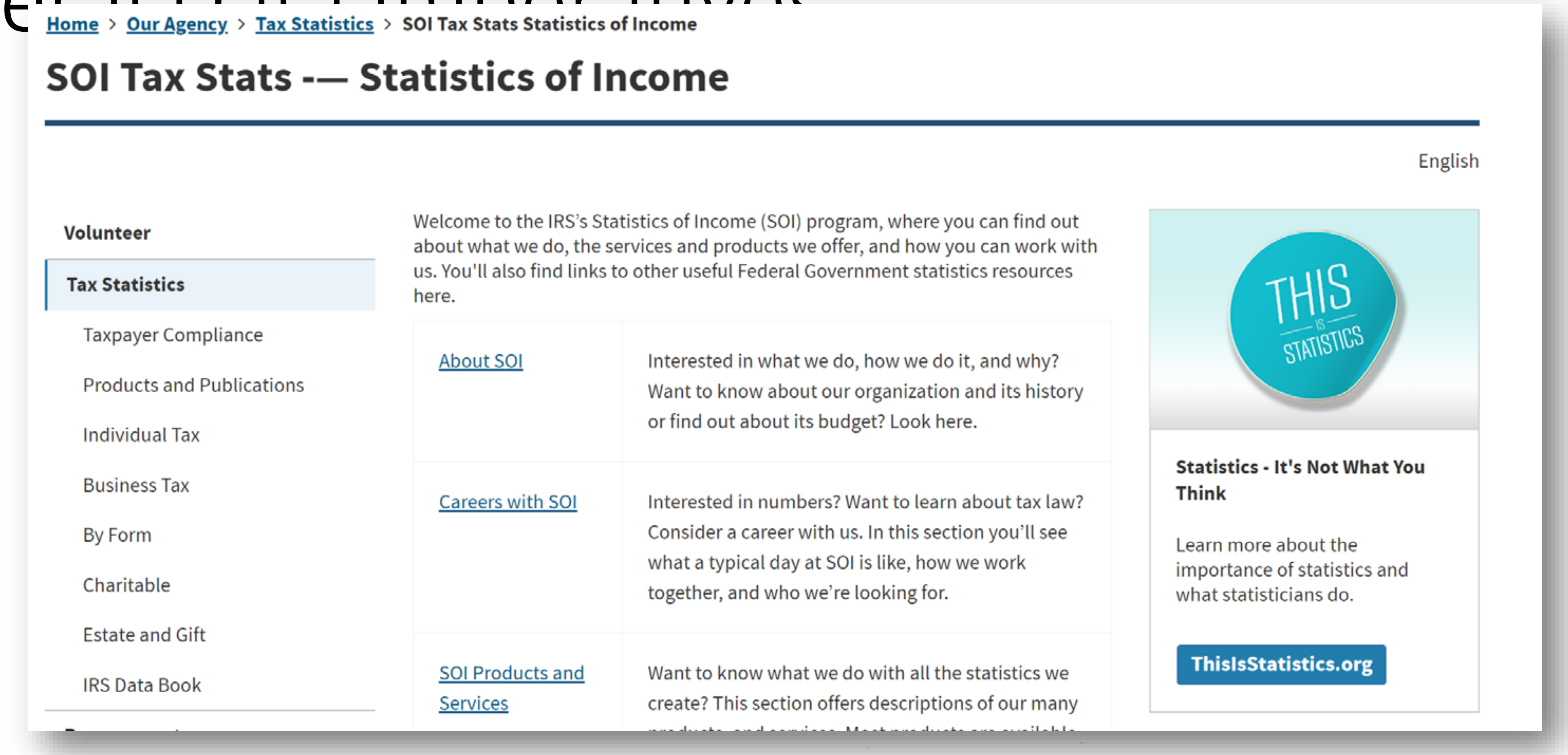
Discussant comments

Julia Lane, NYU

Federal Data Imperatives



Home > Our Agency > Tax Statistics > SOI Tax Stats Statistics of Income

SOI Tax Stats -— Statistics of Income

English

- Volunteer
- Tax Statistics
  - Taxpayer Compliance
  - Products and Publications
  - Individual Tax
  - Business Tax
  - By Form
  - Charitable
  - Estate and Gift
  - IRS Data Book

Welcome to the IRS's Statistics of Income (SOI) program, where you can find out about what we do, the services and products we offer, and how you can work with us. You'll also find links to other useful Federal Government statistics resources here.

About SOI	Interested in what we do, how we do it, and why? Want to know about our organization and its history or find out about its budget? Look here.
Careers with SOI	Interested in numbers? Want to learn about tax law? Consider a career with us. In this section you'll see what a typical day at SOI is like, how we work together, and who we're looking for.
SOI Products and Services	Want to know what we do with all the statistics we create? This section offers descriptions of our many products and services. Most products are available

Statistics - It's Not What You Think

Learn more about the importance of statistics and what statisticians do.

ThisIsStatistics.org

- *Building a Culture that Values Data and Promotes Public Use (practices 1-10)*
- *Governing, Managing, and Protecting Data (practices 11-26)*
- *Promoting Efficient and Appropriate Data Use (practices 27-39)*

Inquiring minds want to know

Privacy vs utility

- What’s the use case?
- Value of approach for protecting privacy
  - Disclosure versus harm
  - What is disclosure (e.g. income of $1,000,000 vs income of $1,000,005
  - Differential impact by groups?
- What is the impact on data quality
  - General estimates
  - Subsamples
  - Outliers

Operational Issues

- Time and Cost?
- How generalizable is approach?
- What is the likely takeup and value to the agency?

What’s the tradeoff?

What’s the use case?

Burman

1. model the effect of current or proposed law on the distribution of tax burdens across income groups, on incentives to work or to make charitable contributions, and on other effects of the income tax.
2. “training dataset”

Feenberg

1. Reproduce the revenue scores of the public use version of the Individual Tax Model. We believe analysis of the ability to reproduce policy-relevant parameters is the most useful way to demonstrate the usability of masked data. Revenue scores are among the most common and important uses of tax micro-data, and lend themselves to this measure
2. Scores are non-linear functions of (typically) subsets of the data

Privacy vs. Utility: Burm

Contribution

1. “meaningful disclosure”
2. Smoothed distribution for outliers

Questions

Protection methodology

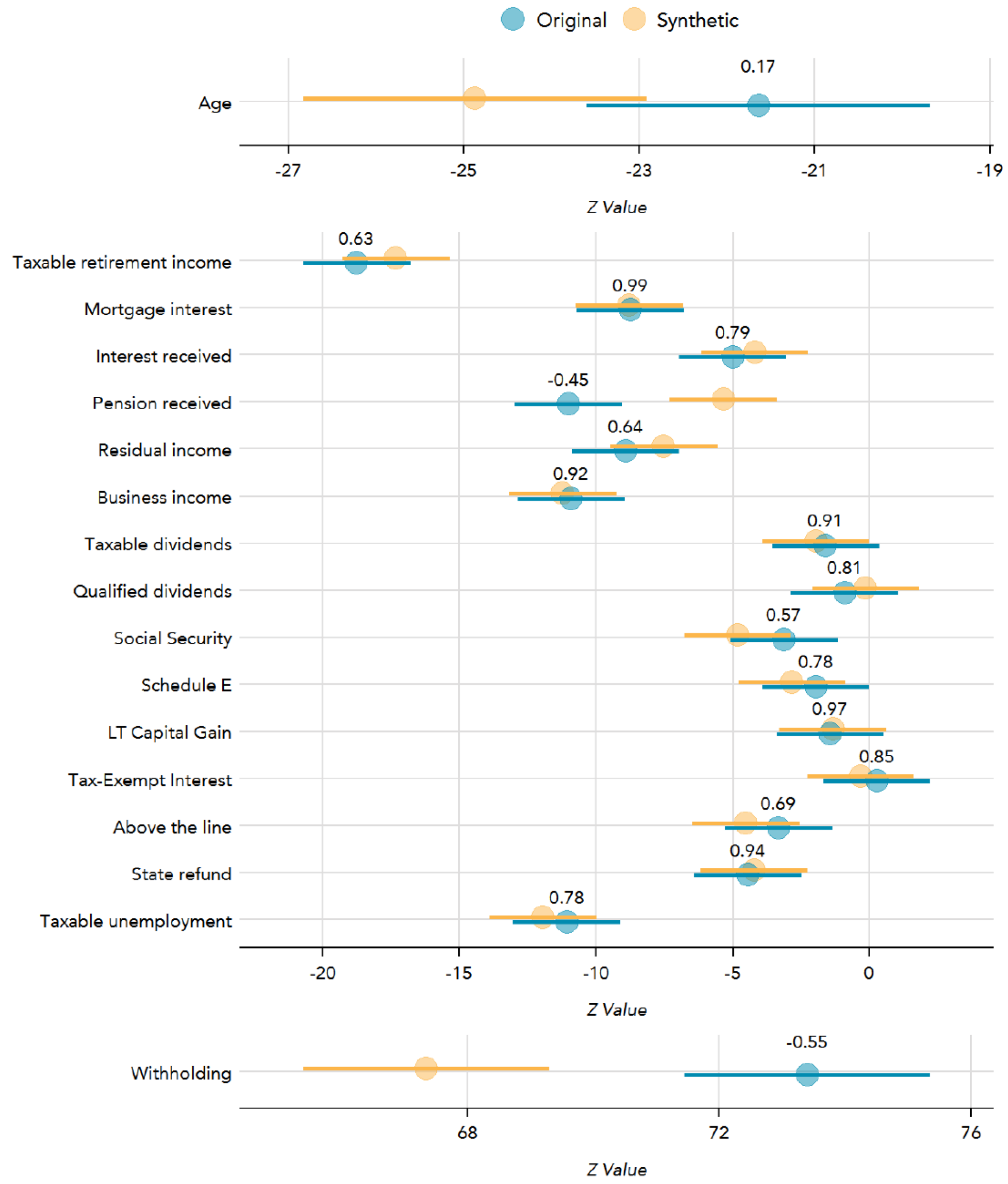
1. Limited set of variables
2. Heavily dependent on sampling
3. Big distance for outliers – but using compressed dist

Quality issues

1. What is the regression model?
2. What’s the loss function?
3. What is “meaningful disclosure”?
4. How will it work with correlations across time?

FIGURE 14

Regression Confidence Interval Overlap



Note: Calculation excludes rows with zeros for all seventeen tax variables.

Privacy vs. Utility: Feenberg

Contribution

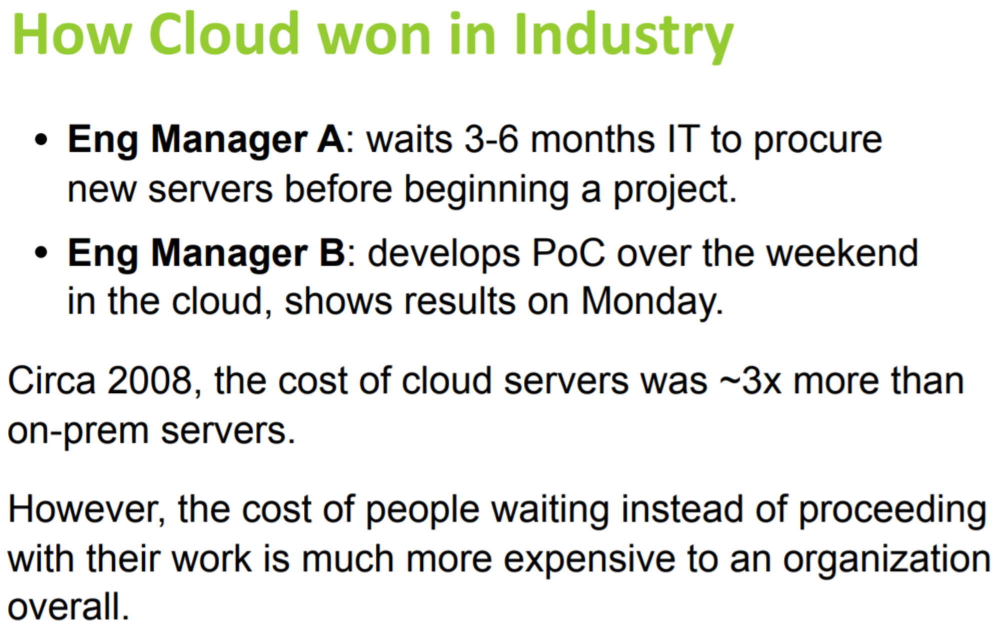
1. Provides clear methodology for examining non-linear functions of small subsets

2. Definition of nearby “twice Euclidan distance of the closest record”

3. Unless the synthetic dataset is explicitly created with a separate distribution for this particular subset of taxpayers, the ability to correct for endogeneity will be lost. It is likely that most potential uses for synthetic data will run into one or more similar roadblocks as a CDF model general enough to anticipate the many possible demands placed on a dataset would be unmanageably complex.

Operational issues

- Time and Cost?
  - Research vs. operations
- How generalizable is approach?
  - Berman – not very
  - Feenberg – not
  - Census experience?
- What is the likely takeup and value to the agency?
  - Berman – validation server needed; value to students



How Cloud won in Industry

- **Eng Manager A**: waits 3-6 months IT to procure new servers before beginning a project.
- **Eng Manager B**: develops PoC over the weekend in the cloud, shows results on Monday.

Circa 2008, the cost of cloud servers was ~3x more than on-prem servers.

However, the cost of people waiting instead of proceeding with their work is much more expensive to an organization overall.

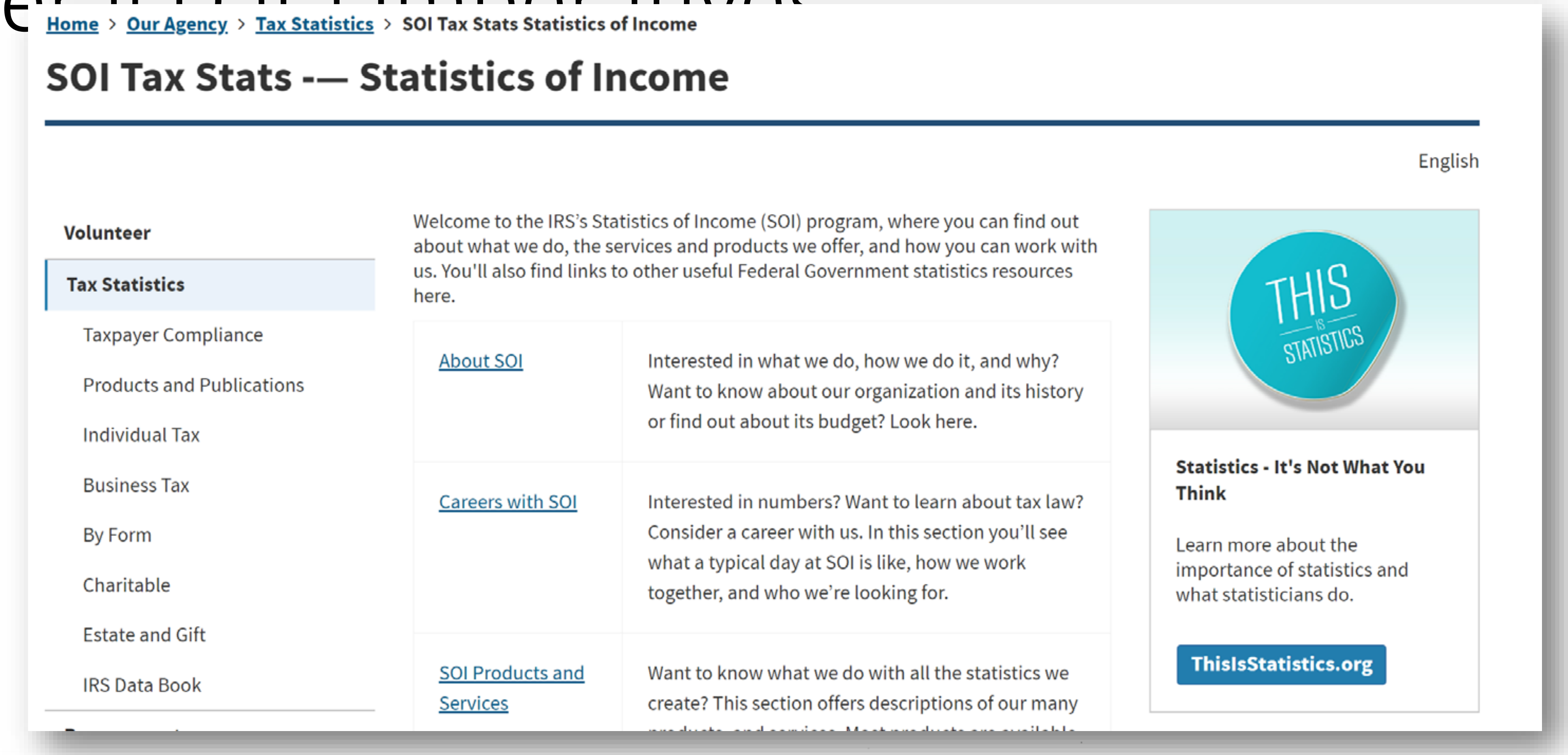
What’s the tradeoff?

What are SOI goals – data cleaning; building expert community; developing new products..

Tiered remote access – with access and utility directly tied

- Safe People
  - Tiers of approved and trained researchers; tiers of legal controls
- Safe Projects
  - Tiers of approved projects, consistent with agency mission and utility
- Safe Settings
  - Tiers of secure environments
- Safe Data
  - Tiers of protected data,
- Safe Outputs
  - Disclosure reviews and export controls

Federal Data Imperatives



Home > Our Agency > Tax Statistics > SOI Tax Stats Statistics of Income

SOI Tax Stats -— Statistics of Income

English

- Volunteer
- Tax Statistics
  - Taxpayer Compliance
  - Products and Publications
  - Individual Tax
  - Business Tax
  - By Form
  - Charitable
  - Estate and Gift
  - IRS Data Book

Welcome to the IRS's Statistics of Income (SOI) program, where you can find out about what we do, the services and products we offer, and how you can work with us. You'll also find links to other useful Federal Government statistics resources here.

About SOI	Interested in what we do, how we do it, and why? Want to know about our organization and its history or find out about its budget? Look here.
Careers with SOI	Interested in numbers? Want to learn about tax law? Consider a career with us. In this section you'll see what a typical day at SOI is like, how we work together, and who we're looking for.
SOI Products and Services	Want to know what we do with all the statistics we create? This section offers descriptions of our many products and services. Most products are available

Statistics - It's Not What You Think

Learn more about the importance of statistics and what statisticians do.

ThisIsStatistics.org

- *Building a Culture that Values Data and Promotes Public Use (practices 1-10)*
- *Governing, Managing, and Protecting Data (practices 11-26)*
- *Promoting Efficient and Appropriate Data Use (practices 27-39)*

Machine Learning: Ann Parker

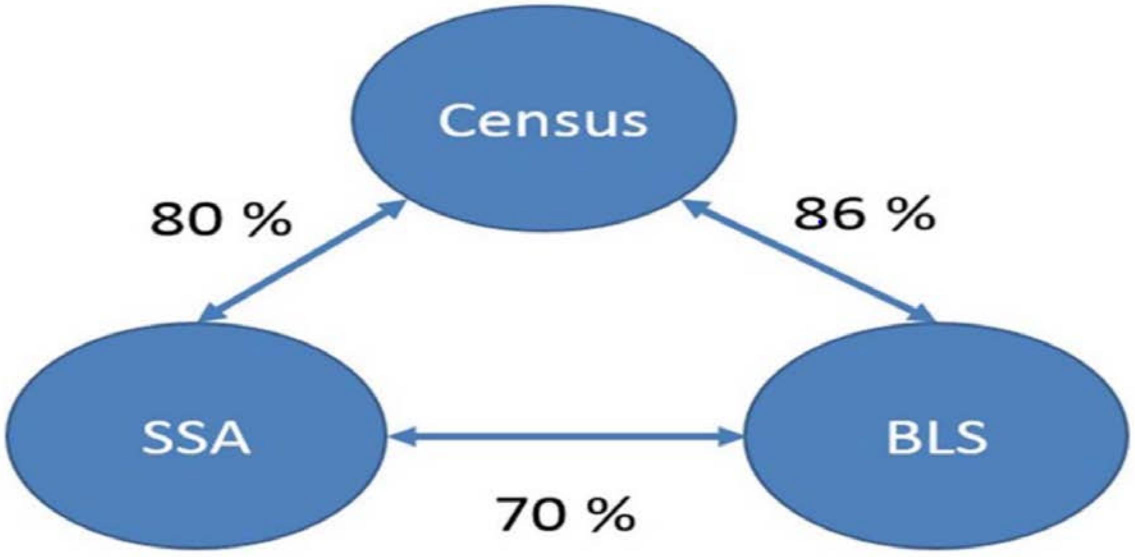
Overall reaction

- Great to see this work
- Careful and thoughtful
- Part of new focus of agencies
  - Using Public Data to Generate Industrial Classification Codes (John Cuffe et al.) – CRIW 2019 (Scraped Google and Yelp)
- Fits in very well with Federal Data Strategy

Technical questions around training data set

- How was training set constructed?
  - Years
  - Sample structure
- Ground truth for training d

Figure 1: Agreement on NAICS Sectors between Census, BLS, and SSA.



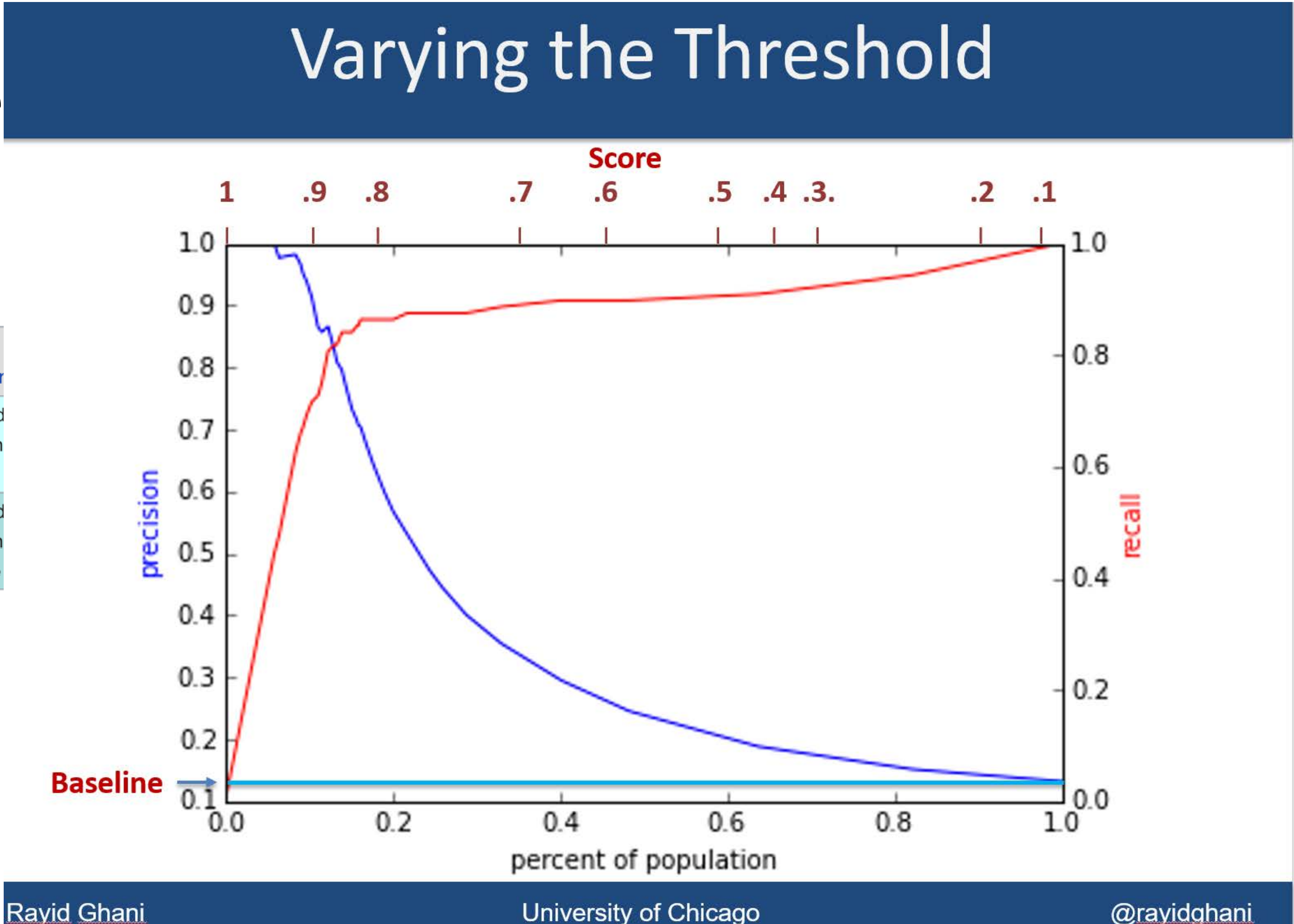
Note: Figure shows the Percentage of BR establishments that share a common 2-digit NAICS sector when present in each respective data source. Source: 2012 Business Register Single Unit Businesses.

Te

Total population

Predicted condition | Predicted condition positive

Predicted condition negative



Varying the Threshold

Rayid Ghani University of Chicago @rayidghani

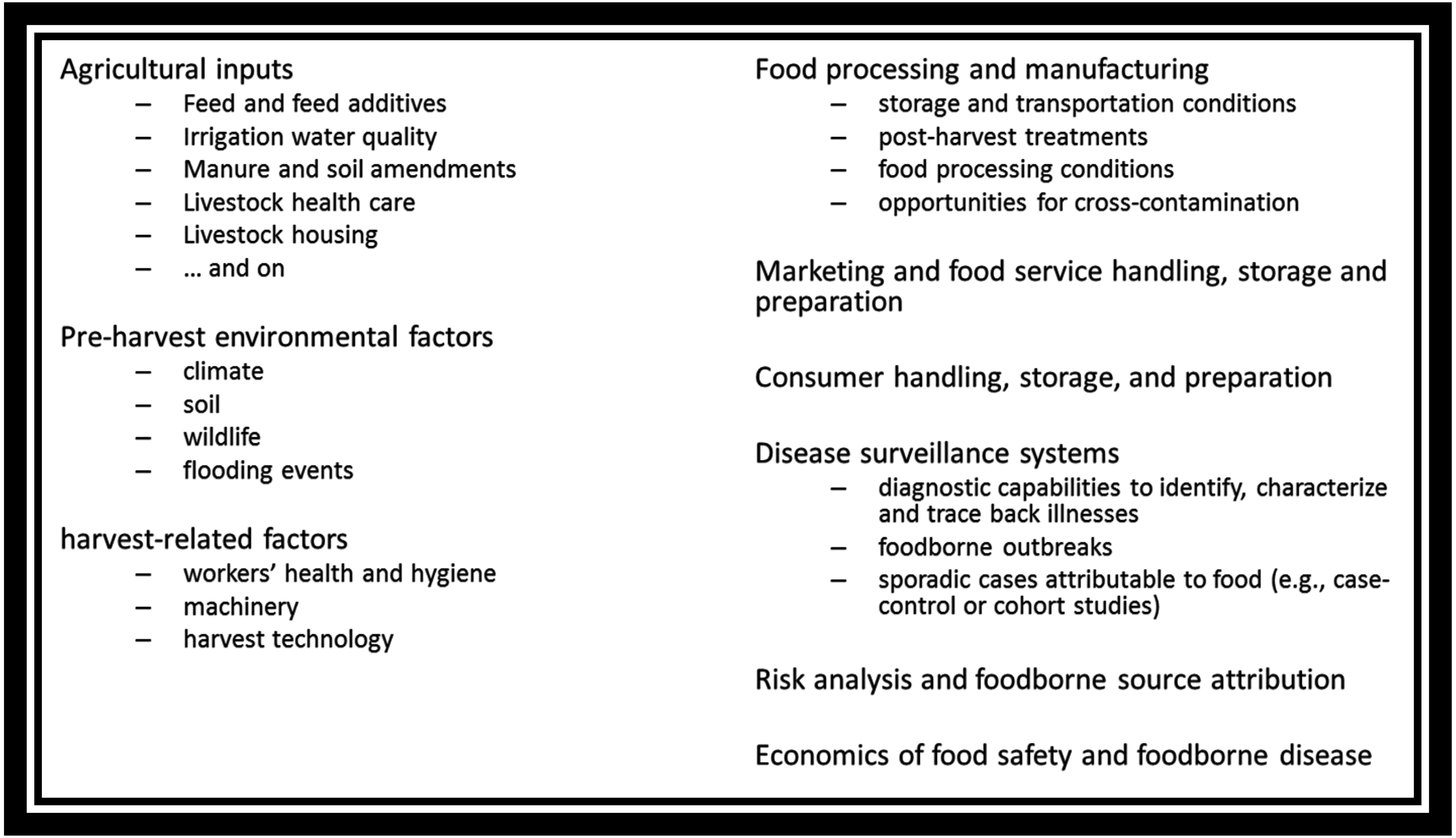
y (ACC) =
+ Σ True negative
population

ry rate (FDR) =
positive
ondition positive

tive value (NPV) =
negative
ndition negative

F_1 score = $\frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}$

Food Safety Research’s Scope Spans the Complete Farm-to-Fork Chain



Agricultural inputs

- Feed and feed additives
- Irrigation water quality
- Manure and soil amendments
- Livestock health care
- Livestock housing
- … and on

Pre-harvest environmental factors

- climate
- soil
- wildlife
- flooding events

harvest-related factors

- workers’ health and hygiene
- machinery
- harvest technology

Food processing and manufacturing

- storage and transportation conditions
- post-harvest treatments
- food processing conditions
- opportunities for cross-contamination

Marketing and food service handling, storage and preparation

Consumer handling, storage, and preparation

Disease surveillance systems

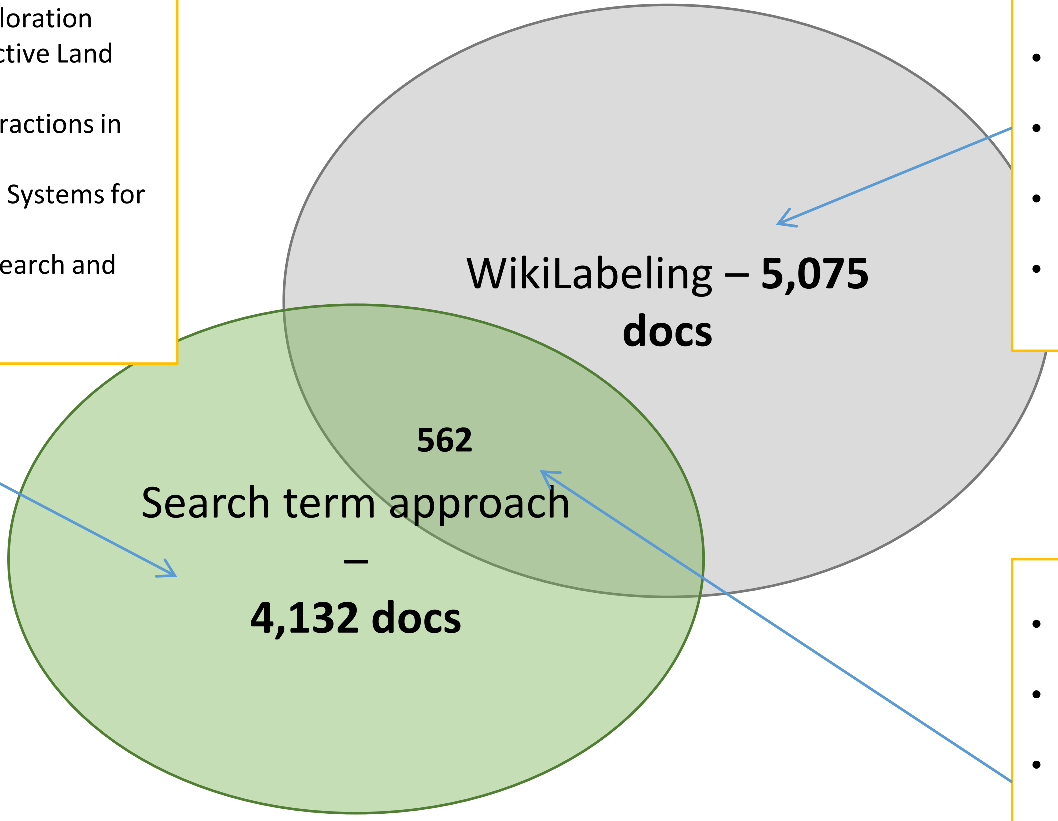
- diagnostic capabilities to identify, characterize and trace back illnesses
- foodborne outbreaks
- sporadic cases attributable to food (e.g., case-control or cohort studies)

Risk analysis and foodborne source attribution

Economics of food safety and foodborne disease

Example: NSF (2000-2013) N=154,925

- BREAD: Improving Water Acquisition in Maize with Root Traits that Reduce the Metabolic Cost of Soil Exploration
- Collaborative Research: A Pilot Project on Interactive Land Use and Climate Predictions
- Molecular and Cellular Biology of Biotrophic Interactions in Rice Blast Disease
- SBIR Phase II: SaaS-Based Procurement and CRM Systems for Local Food Markets
- NIRT: Building Capacity for Social and Ethical Research and Education in Agrifood Nanotechnology


WikiLabeling – **5,075 docs**
562
Search term approach – **4,132 docs**

- BE/GEN-EN: Development of Methods Linking Genomic and Ecological Responses in a Freshwater Sentinel Species
- Environmental Implications of Engineered Nanomaterials on the Important Environmental Model Daphnia
- Workshop on Modeling the Rapid Evolution of Infectious Diseases
- The Genetic Architecture of Adaptation in Laboratory Yeast Populations
- Cell-to-Cell Signaling in E. Coli

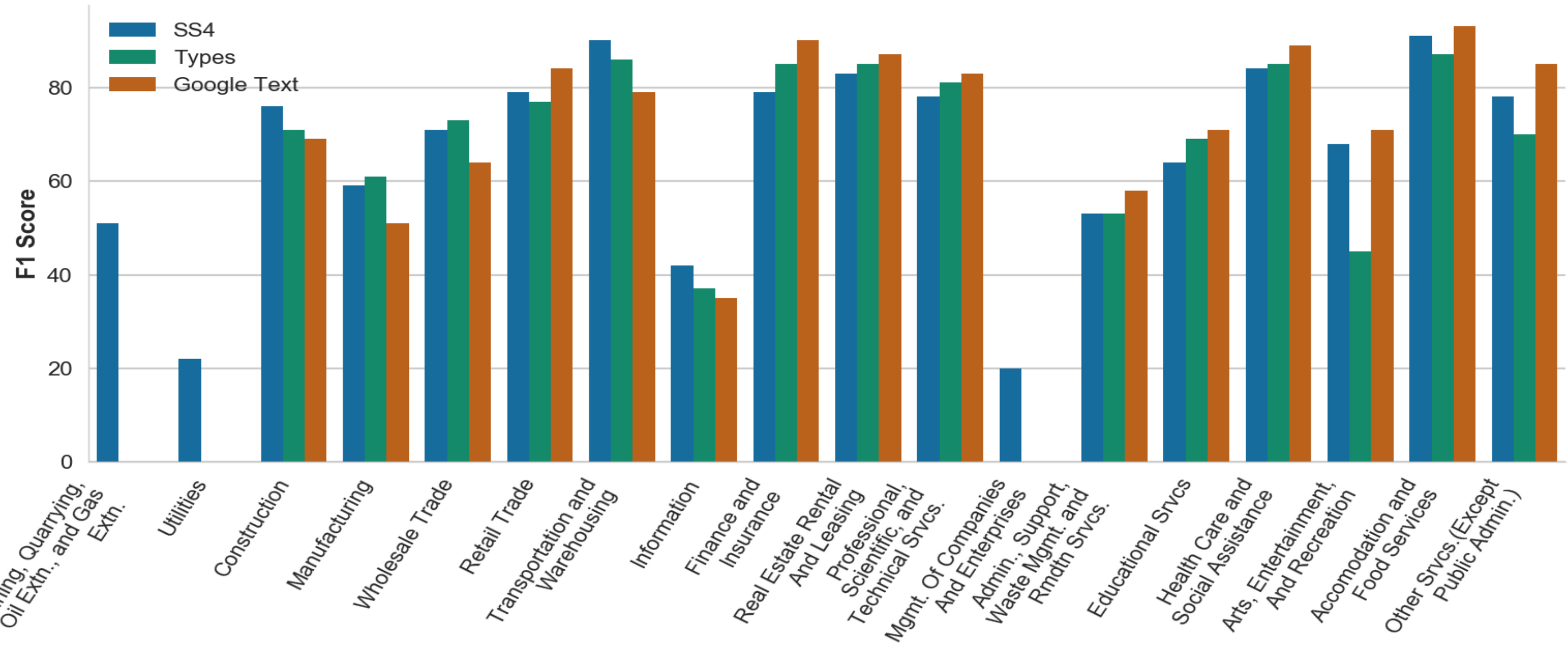
- SBIR Phase II: A Microfluidic-Based Biosensor for Food Pathogen Detection
- STTR Phase I: Engineered Bacteriophage Based Biosensor for Rapid Detection of Viable Foodborne Pathogens
- Biotechnology and Bioinformatics Program: The Institute For Food Safety
- Workshop on Genomic Analysis of Plant-Associated Microorganisms; Washington, D.C., April 9 - 11, 2002
- Dosage dependent genes affecting seed composition and weight

Common Words in Some Industries

Industry	Most Common Words in Google	Most Common Words in SS4
Information	new, 2018, business, news, contact	media, internet, production, publishing, software
Health Care And Social Assistance	doctor, care, staff, office, time	care, therapy, medical, physical, doctor
Retail Trade	store, staff, friendly, time, best	store, sales, convenience, retail, internet
Accomodation And Food Services	food, staff, friendly, time, best	restaurant, food, bar, shop, retail
Mining, Quarrying, Oil Extn., And Gas Extn.	safety, contact, oil, energy, construction	oil, gas, oilfield, drilling, energy
Finance And Insurance	insurance, financial, business, investment, help	investment, insurance, management, financial, advice

Using Public Data to Generate Industrial Classification Codes (John Cuffe et al.

Results: Model Performance



Source: Using Public Data to Generate Industrial Classification Codes (John Cuffe et al.) – CRIW 2019