



Session 4. Doing More With Less

Moderator:

Tom Hertz

IRS: RAAS

Can Machine Learning Improve Correspondence Audit Case Selection? Considerations for Algorithm Selection, Validation, and Experimentation

Ben Howard

MITRE Corporation

Audit Productivity, Taxpayer Service, and Compliance: Can a Service Mindset Overcome a Dwindling Enforcement Budget?

Mary Marshall

Louisiana Tech Univ.

Using the Internal Revenue Service Program Assessment Model Optimizer to Inform Resource Allocation Decisions

Clay Swanson

IRS: SB/SE

Discussant:

Arnie Greenland

A. G. Analytics, LLC

Can Machine Learning Improve Correspondence Audit Case Selection?

Considerations for Algorithm Selection, Validation, and Experimentation

BEN HOWARD¹, LUCIA LYKKE¹, DAVID PINSKI¹, AND ALAN PLUMLEY²

¹The MITRE Corporation

²Internal Revenue Service

June 18, 2020



Internal Revenue
Service

MITRE

SOLVING PROBLEMS
FOR A SAFER WORLD™

Why machine learning for correspondence audits?

Correspondence audits are conducted primarily by mail. Each category of correspondence audit focuses on a narrowly defined groups of taxpayers and examines a small number of line items.

Traditional methods for selecting taxpayers to audit rely on rules of thumb and few data points.



Machine learning has two advantages:

1. Exploit many data points such as reporting history from past returns.
2. Detect patterns among data points in order to predict an outcome.



Definitions of Terms

Revenue (Continuous): The assessed tax amount for an audited return. Revenue may or may not be collected from the taxpayer.

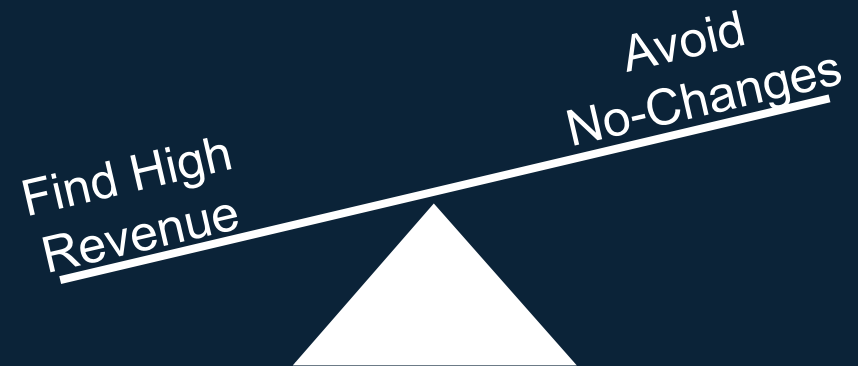
No-Changes (Binary): Audits that result in no assessed tax change for a return (i.e., a “false positive”). These audits are not cost effective for the IRS and result in taxpayer burden.

Trade-offs in Modeling Approach

In a **perfect world**, avoiding no-changes would result in higher revenue.

Prediction is **always imperfect** and trying to avoid no-changes can result in missing higher revenue.

For example, more complex returns (**more uncertainty**) might have more noncompliance issues (**higher revenue**) but can also be harder to detect before an audit.



Background

Data mining techniques at the IRS

- DIF scores have long been used to predict noncompliance and select returns for field audits (Brown and Mazur 2003)
- Based on **full return** audits representing the full individual taxpayer population; this does not translate to correspondence audits, which look at **limited line items** for narrow slivers of the taxpayer population

Other applications of data mining techniques to audit selection

Pilot studies show increased revenue and ability to detect noncompliance:

- Minnesota Sales and Use Tax (Hsu *et al.* 2015)
- Texas Sales Tax (Micci and Barrecca 2009)
- Taxpayers in Ireland (Cleary *et al.* 2011)

Review of Past Correspondence Audit Experiments

For several years, MITRE and RAAS have worked with Campus Exam on operational **random control trials** to test the effectiveness of machine learning models for two different categories of audit.

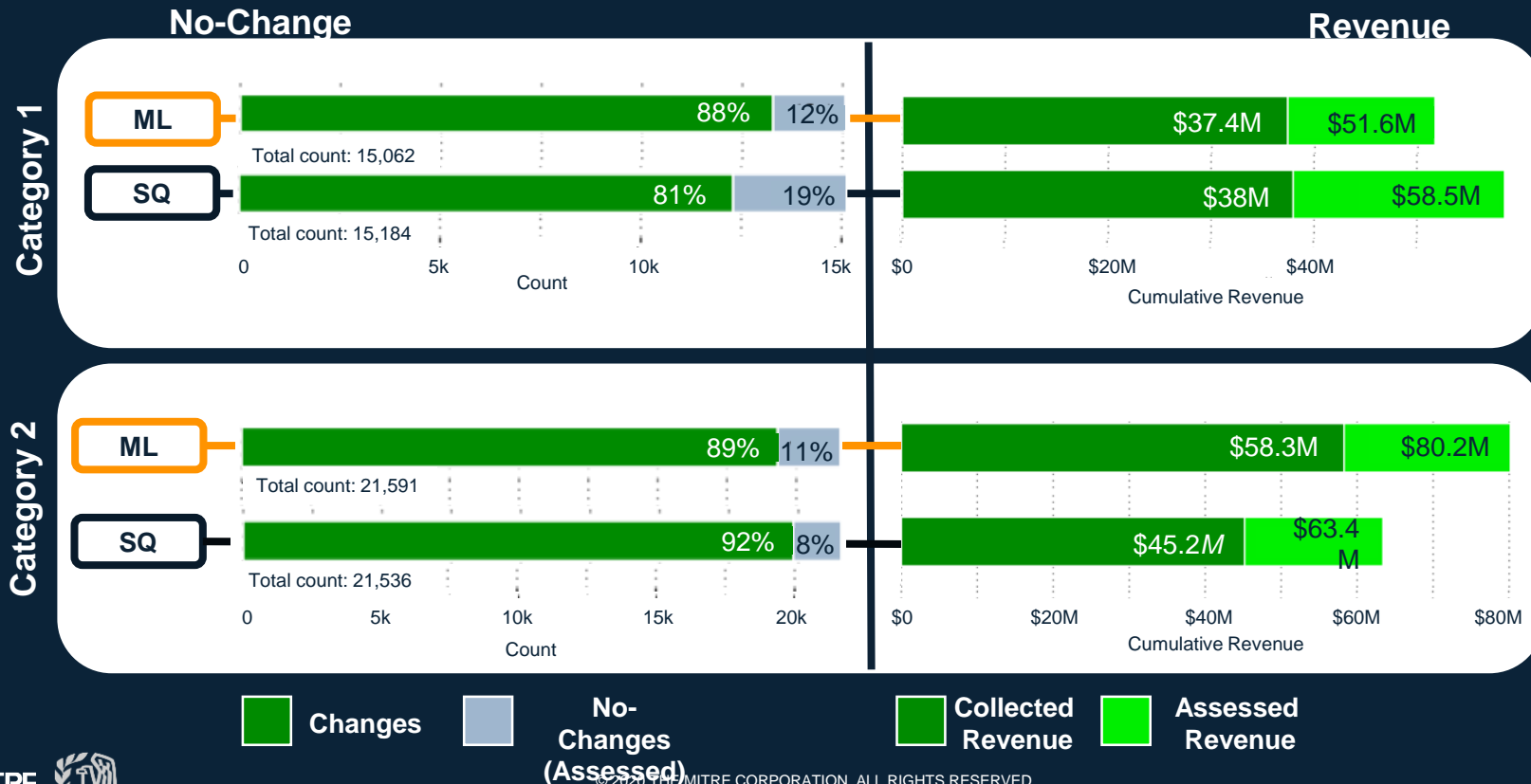
We examine both outcomes of interest for the experiments:

1. Revenue (assessed revenue and collected revenue)
2. No-change rates (“false positives”)

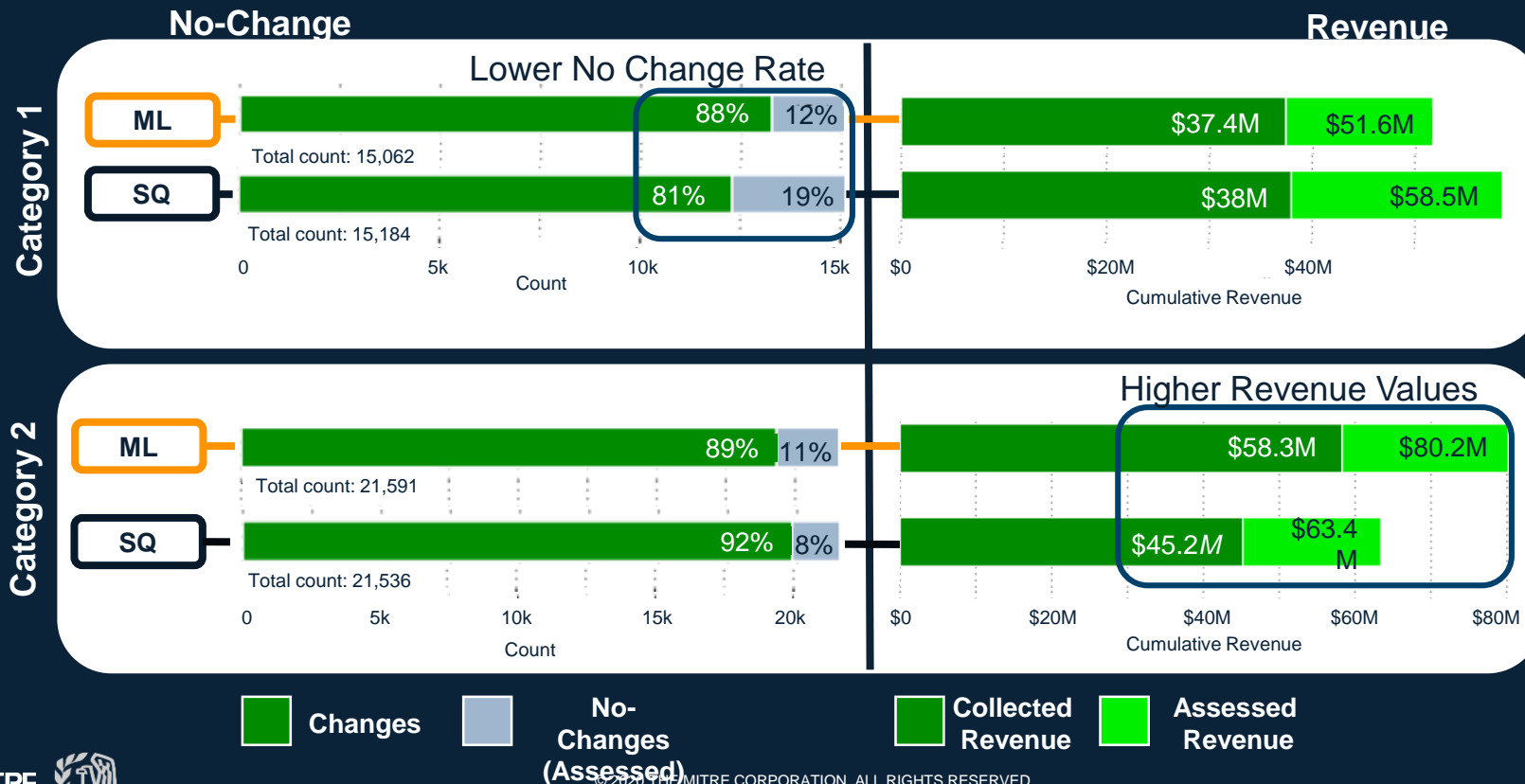
A different algorithm was applied to each audit category and each shows different kinds of improvement from the experiment:

- Category 1 - **Lower no-change rate** & equal collected revenue
- Category 2 – Much **higher revenue** but slightly higher no-change rate

Comparison of Tax Year 2014 Experiments



Comparison of Tax Year 2014 Experiments



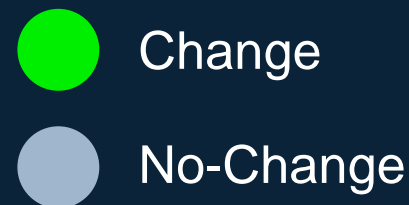
Considering Different Strategies

Learning from Past Experiments and
Refining Methodology

Three Strategies Applied to Audit Category 1

Pairwise Ranking

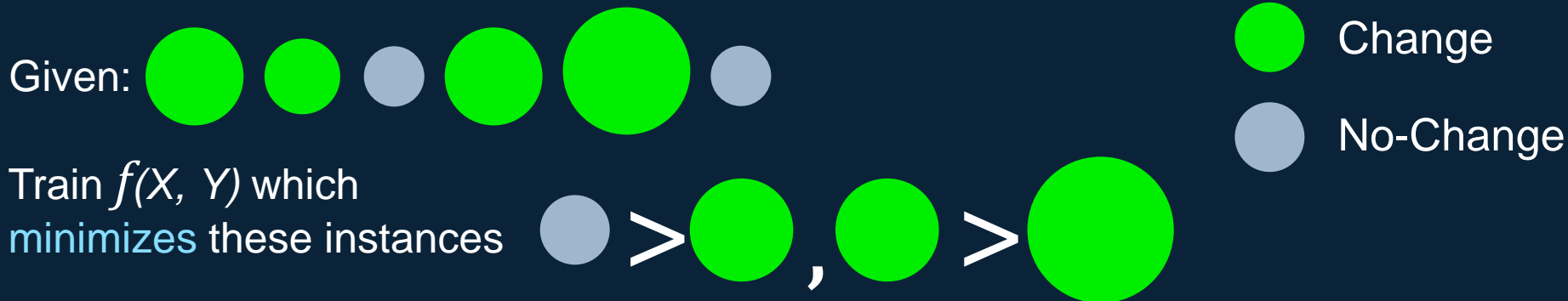
- Train a greater than / less than function on historical audit data.
- Training objective is to minimize the number of inversions in ranking - this occurs when any given pair of results is in the wrong order.



Three Strategies Applied to Audit Category 1

Pairwise Ranking

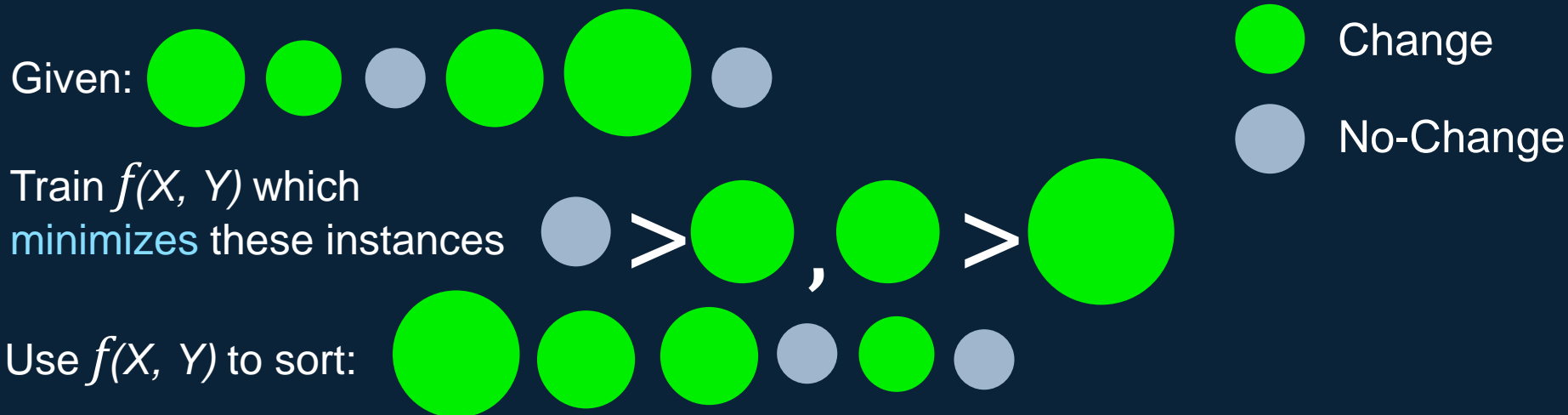
- Train a greater than / less than function on historical audit data.
- Training objective is to minimize the number of inversions in ranking - this occurs when any given pair of results is in the wrong order.



Three Strategies Applied to Audit Category 1

Pairwise Ranking

- Train a greater than / less than function on historical audit data.
- Training objective is to minimize the number of inversions in ranking - this occurs when any given pair of results is in the wrong order.

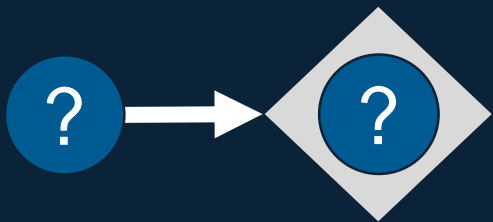


Three Strategies Applied to Audit Category 1

Two-Stage Regression

1. Predict if the return is a likely no-change.
2. If it is a predicted change, predict the revenue. Otherwise do not select.

Given an
unlabeled
return



Is this a
predicted
no-change?



No-Change
Model

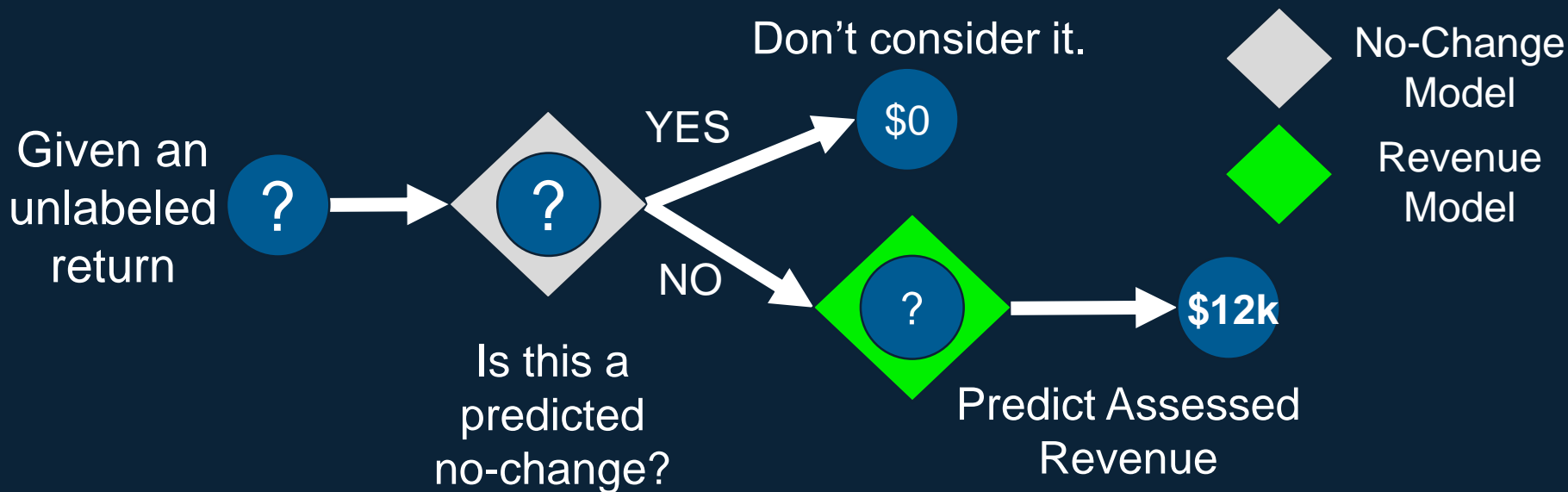


Revenue
Model

Three Strategies Applied to Audit Category 1

Two-Stage Regression

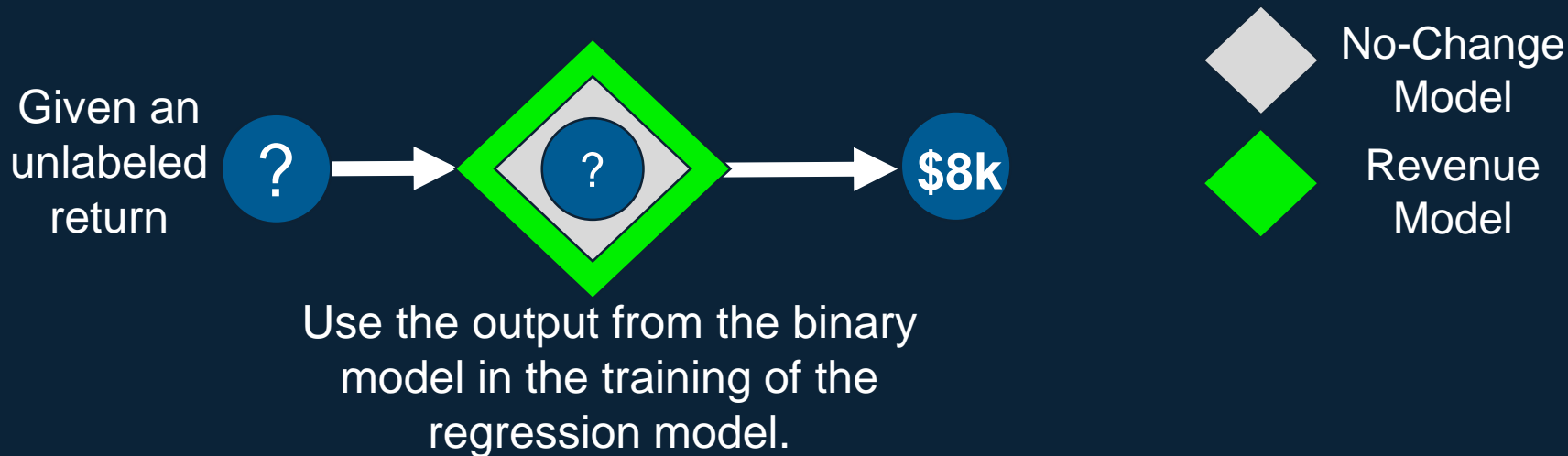
1. Predict if the return is a likely no-change.
2. If it is a predicted change, predict the revenue. Otherwise do not select.



Three Strategies Applied to Audit Category 1

Penalized Regression

The predicted probability of no-change is used to penalize predicted revenue during the training of the regression model.



Validating Expected Results for Audit Category 1

Comparing to Status Quo Method and
Perfect Knowledge (Retrospective Ranking)

Validation Overview

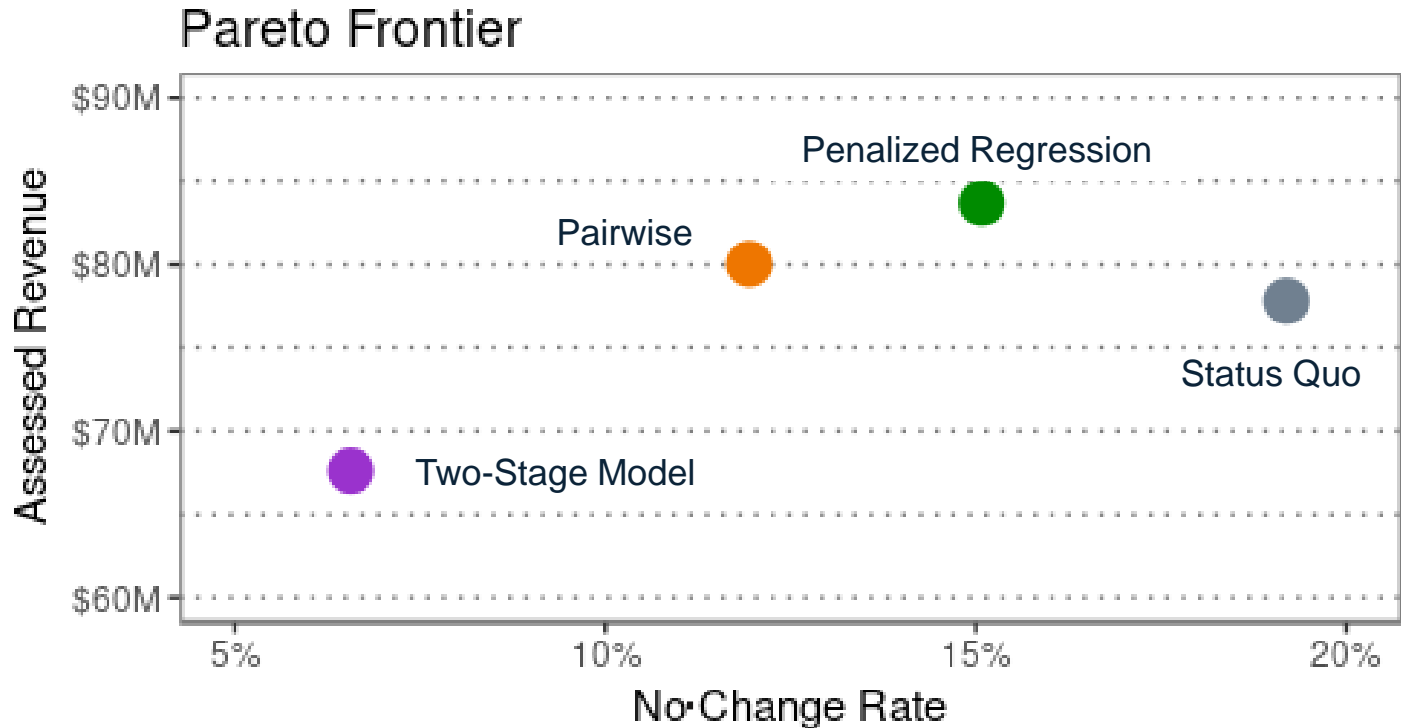
We use **Gradient Boosted Machines** (GBMs) with the `xgboost` package (Chen and Guestrin, 2016).

Models are initially developed on older tax years and out-of-time validation is used on the most recent year of audit records.

Perfect Knowledge Ranking: The actual results of the audit. This is the `best possible` ranking a dataset can give us. We compare all other methods to this one.

Pareto Frontier

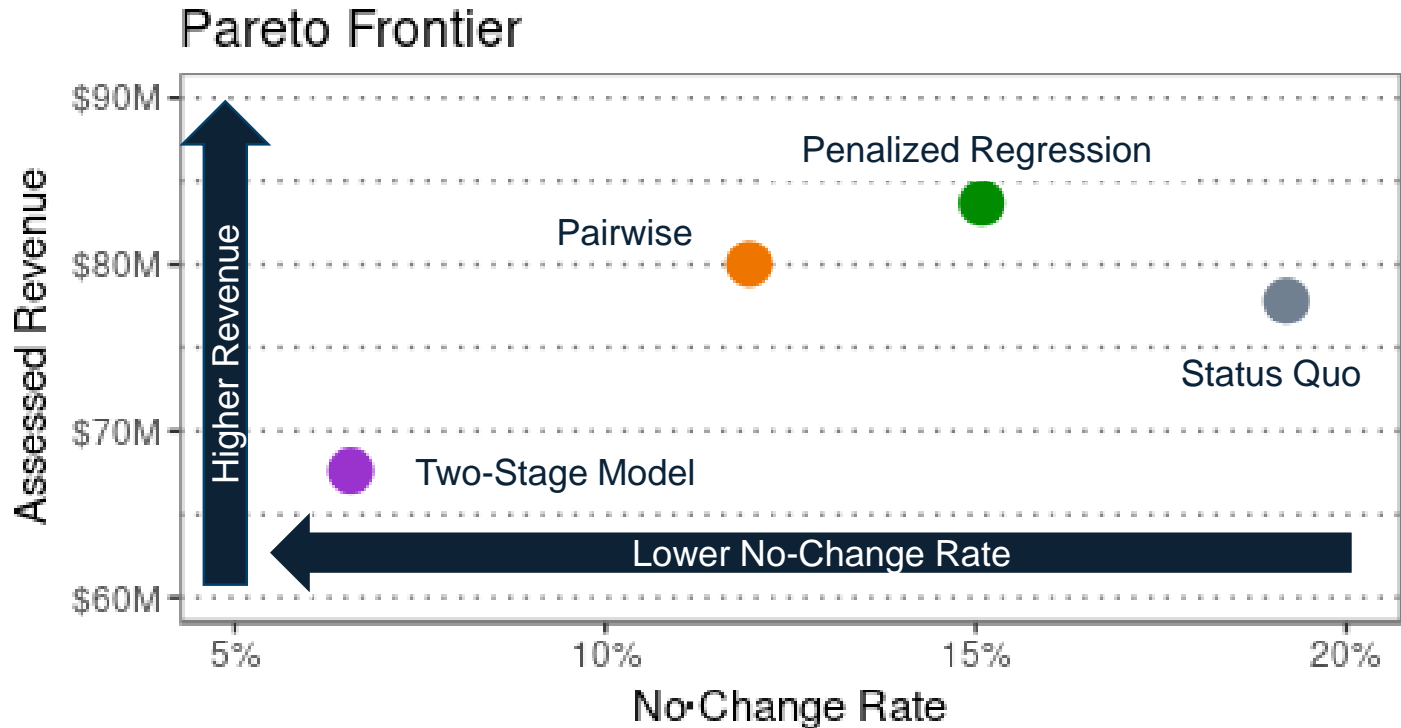
Plots cumulative revenue against no-change rate at 50% of inventory



The Pareto Frontier captures a snapshot of the rankings to visually evaluate the trade-off.

Pareto Frontier

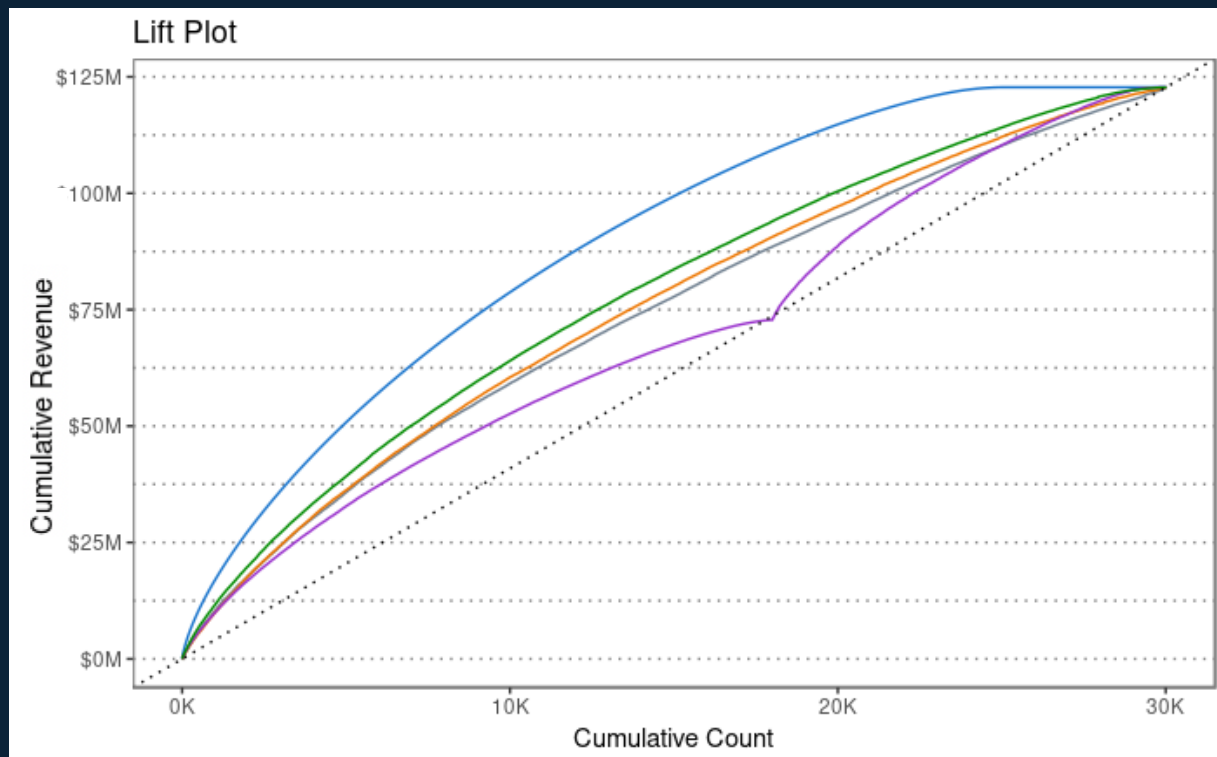
Plots cumulative revenue against no-change rate at 50% of inventory



- The Two-Stage Model is excellent for reducing the no-change rate but is poor at maximizing revenue.
- Penalized Regression and Pairwise perform better on both objectives compared to Status Quo.

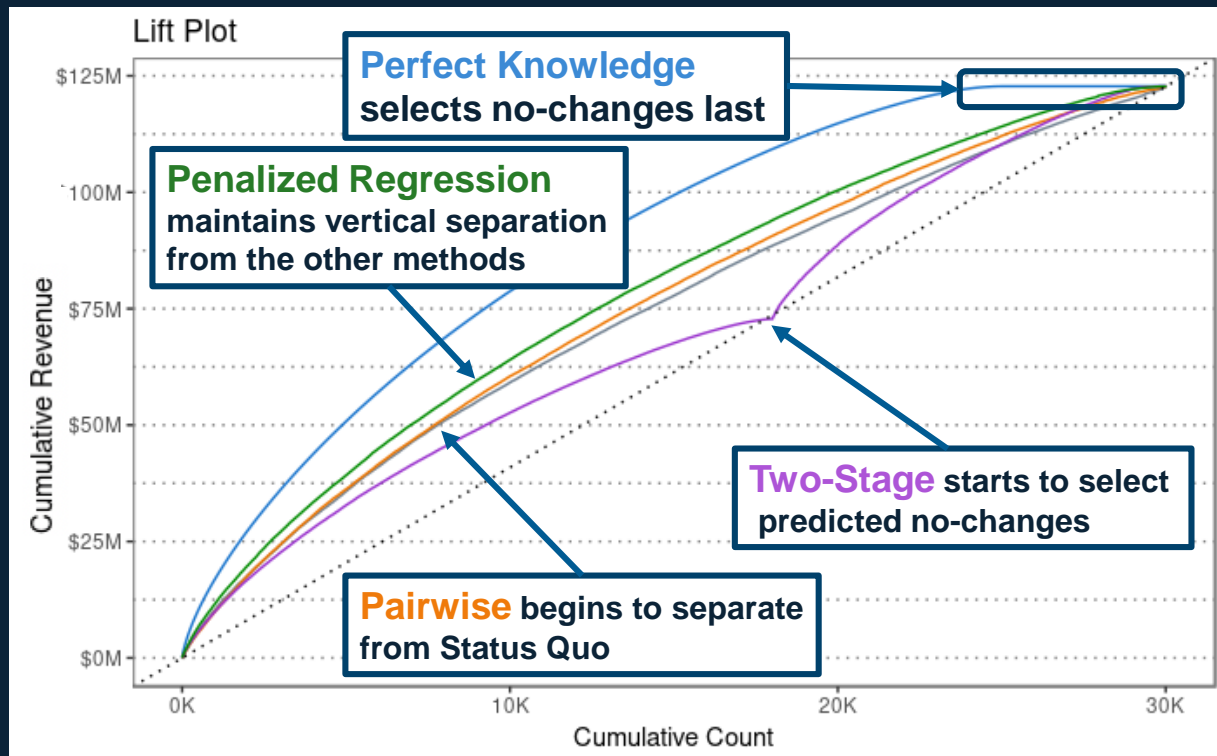
Lift Plot

Plots predicted cumulative revenue by priority rank



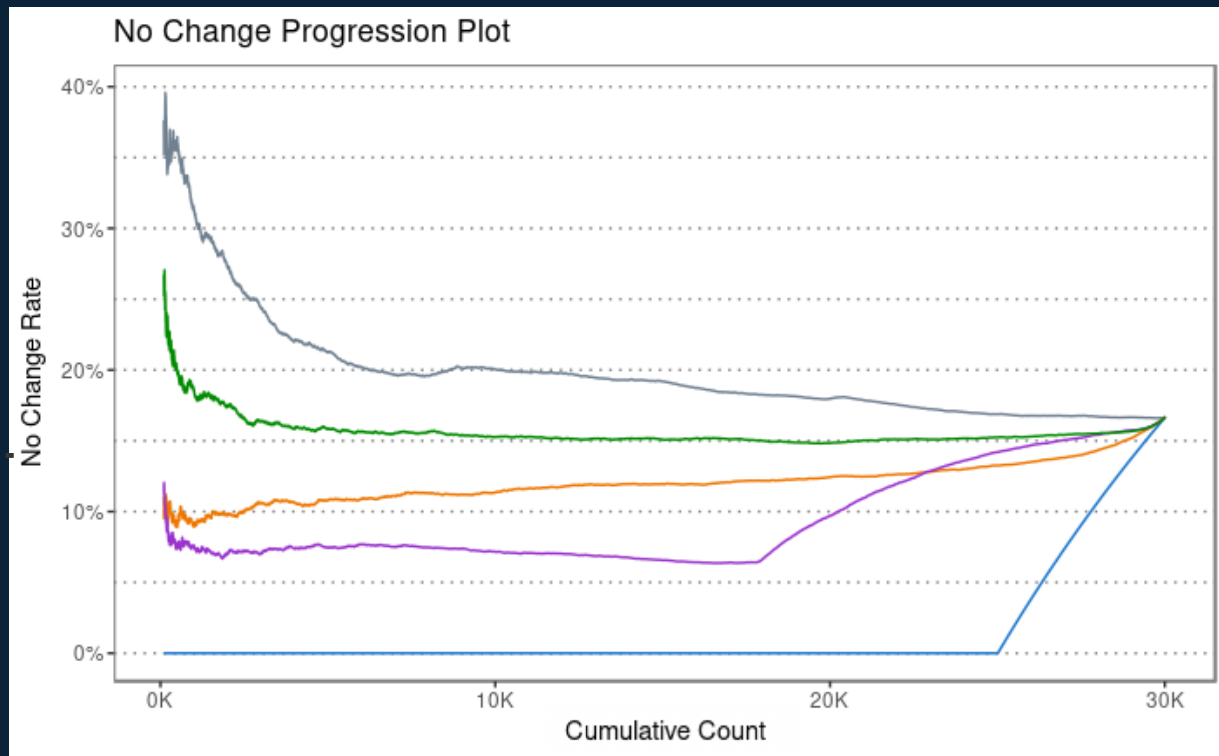
Lift Plot

Plots predicted cumulative revenue by priority rank



No-Change Progression

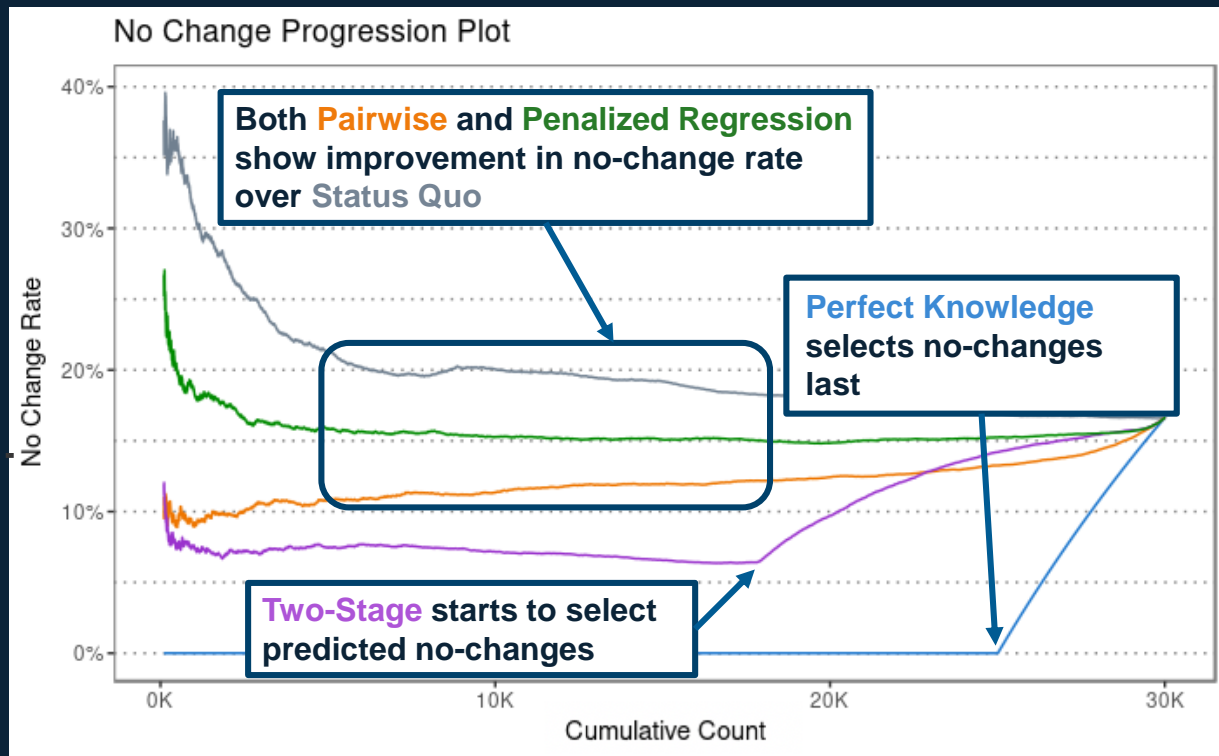
Shows no-change rate by priority rank



- Perfect Knowledge
- Penalized Regression
- Pairwise
- Two-Stage Model
- Status Quo

No-Change Progression

Shows no-change rate by priority rank



Discussion

- The strategy used to balance these two objectives is only one consideration in developing machine learning models, there are **many considerations** that impact prediction performance. (e.g. Quality data and features, what algorithm and package to use, etc...)
- There are more considerations for **decision makers** other than maximizing revenue and reducing no-change rate.
- Machine Learning will be the right tool in **some** situations but **not** others.

References

- Brown, Robert E. and Mazur, Mark J. [IRS's Comprehensive Approach to Compliance Measurement](#), *National Tax Journal*, 56:3, pp. 689-700. 2003.
- Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785-794. 2016.
- Cleary, Duncan, and Revenue Irish Tax. "Predictive analytics in the public sector: Using data mining to assist better target selection for audit." In *ECEG2011-Proceedings of the 11th European Conference on EGovernment: ECEG2011*, p. 168. Academic Conferences Limited, 2011.
- Hsu, Kuo-Wei, Nishith Pathak, Jaideep Srivastava, Greg Tschida, and Eric Bjorklund. "Data mining based tax audit selection: a case study of a pilot project at the Minnesota department of revenue." In *Real world data mining applications*, pp. 221-245. Springer, Cham, 2015.
- Micci-Barreca, Daniele, and Satheesh Ramachandran. "Improving tax administration with data mining." *White paper. Elite Analytics LLC* (2004).

Contact us:

Ben Howard bhoward@mitre.org

Lucia Lykke Ph.D. llykke@mitre.org

David Pinski dpinski@mitre.org

Alan Plumley, Ph.D. Alan.Plumley@irs.gov



**Internal Revenue
Service**

MITRE

**SOLVING PROBLEMS
FOR A SAFER WORLD™**

Audit Productivity, Taxpayer Service, and Compliance: Can a Service Mindset Overcome a Dwindling Enforcement Budget?

Nina S. Collum

Assistant Professor

Louisiana Tech University

Susan Journey

Associate Professor

Oklahoma City University

Mary E. Marshall

Assistant Professor

Louisiana Tech University

IRS/TPC Joint Research Conference on Tax Administration

June 18, 2020

Washington, D.C.



- **Continuously declining resources for tax authorities across jurisdictions, resulting in:**
 - **22% fewer IRS employees (TRAC 2019)**
 - **Limited ability to upgrade “antiquated” systems (National Taxpayer Advocate 2019)**
 - **An increased workload (NTA 2019)**
 - **Overworked employees with low morale (NTA 2018)**





- **Audit Productivity**

- The amount of noncompliance detected by a tax authority during an investigation of the return.

- **Research Questions**

- How does declining audit productivity influence subsequent compliance behavior?
- If so, can we address the reduced compliance with other, less costly efforts and initiatives with a taxpayer service approach?

- Conventional models of tax evasion incorporate the anticipation of an audit

$$EU = (1 - p) U(v + t(y - x)) - p U(v - \theta(y - x)),$$

Detection Probability → Audit Penalty

- Decreased audit productivity likely reduces perceived detection probability.

H1: Taxpayers who experience partially-productive audits are less likely to increase compliance in subsequent years than those who experience fully-productive audits.





- Taxpayers who receive service from the tax authority are more likely to file a tax return and to pay more of their liability. (Alm *et al.* 2010)
- The mere presence of an information service reduces negative reactions to being audited, even if the service is not accessed. (Vossler and Gilpatrick 2018)
- We focus on highlighting the existing service element of the IRS rather than creating a new (probably costly) initiative.

H2: Taxpayers who experience partially-productive audits are more likely to increase compliance in subsequent years when they also experience increased service efforts.

- **Manipulated Independent Variables:**
 - Audit Productivity (100% or 50%)
 - Message (None, Bill of Rights, Assistance)
- **Three experimental rounds**
 - Simulated earnings and reporting tasks
 - First round: Practice
 - Second round: Simulated audit
 - Third round: Service message
- **Dependent Variable:**
 - Change in compliance from round 2 (audited round) to round 3 (message round)

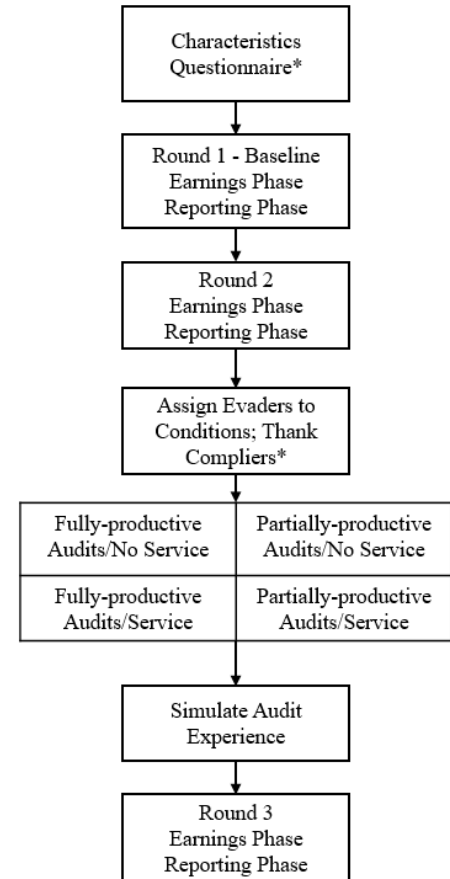


TABLE 2
Analysis of H1

Panel A. Mean [Std. Dev.] of Income Reported (n=271)

<u>Condition</u>	<u>Round 1</u>	<u>Round 2</u>	<u>Round 3</u>	<u>Change</u>
Full/No Message (n=46)	3,935 [555]	3,500 [491]	4,413 [574]	913 [508]
Full/Message (n=88)	4,648 [401]	4,170 [355]	5,625 [415]	1,455 [368]
Partial/No Message (n=46)	3,826 [555]	3,717 [491]	2,326 [574]	(1,391) [509]
Partial/Message (n=91)	4,022 [395]	4,088 [349]	4,517 [408]	429 [362]

Income reported is the average amount of income in Lira (the experimental currency) reported by participants in a given condition for each round. The most participants could report in a given round is 10,000 Lira.

H1: Full/No Message > Partial/No Message
 $t - \text{stat} = 3.417; p = 0.031$

Messages



Congress passes the tax laws and requires taxpayers to comply; however, the IRS is responsible for enforcing those laws. Thus, the IRS mission is to provide America's taxpayers top quality service by helping them understand and meet their tax responsibilities and enforce the law with integrity and fairness to all.

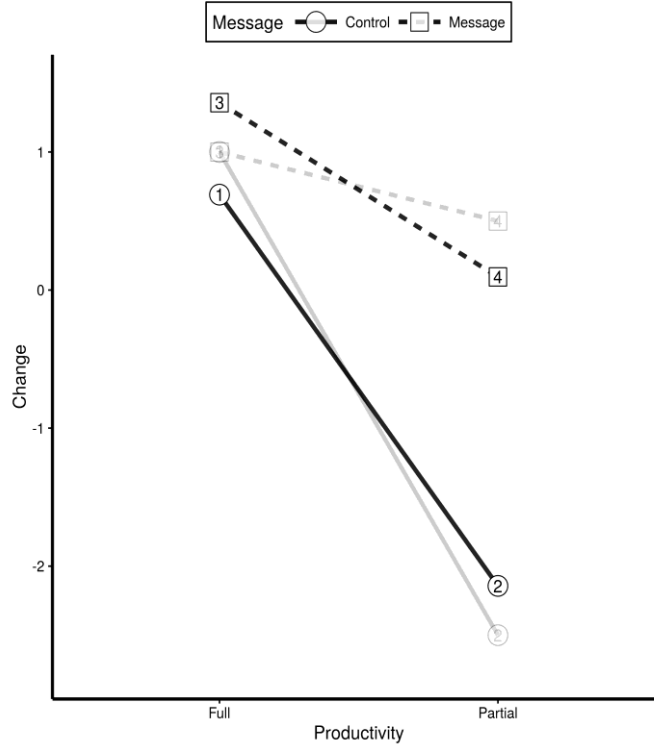
The IRS wants to make it easier for you to make a complete and accurate return. We are here to give you advice and support if you need it.

The IRS mission is to provide America's taxpayers top quality service by helping them understand and meet their tax responsibilities and enforce the law with integrity and fairness to all.

In order to fulfill the mission, the IRS has outlined your rights during the process in a document entitled Taxpayer Bill of Rights, which assures taxpayers are offered the right to:

1. Be informed
2. Quality service
3. Pay no more than the correct amount of tax
4. Challenge the IRS's position and be heard
5. Appeal an IRS decision in an independent forum
6. Finality
7. Privacy
8. Confidentiality
9. Retain representation, and
10. A fair and just tax system.

Results – H2



<u>Comparison</u>	<u>t-statistic</u>	<u>df</u>	<u>p-value</u>
Full Control v. Partial Control	3.277	267	0.001
Full Control v. Full Message	-0.882	267	0.378
Full Control v. Partial Message	0.794	267	0.428
Partial Control v. Full Message	-4.639	267	<0.001
Partial Control v. Partial Message	-2.983	267	0.003
Full Message v. Partial Message	2.035	267	0.042



Process Measure of Audit Productivity Manipulation

- When asked whether “**the auditor did a good job identifying any income I omitted from my tax return,**” participants in the fully productive audit condition reported significantly higher agreement than those in the partially-productive audits. ($p < 0.001$)
- Participants who experienced a fully-productive audit were more likely to report a focus on “**fulfilling obligations to society**” than “paying as little tax as possible” than those in partially-productive audits. ($p = 0.034$)



Process Measures for Service Reminder Message Manipulation

- Responses to dependent variables and post-experimental questions were not different across the two messages. We collapse for brevity.
- Participants who viewed a message were:
 - More likely to indicate the **IRS was focused on providing service (rather than punishing criminals)** than those who did not view a message ($p < 0.001$)
 - More likely to report a focus on “**fulfilling obligations to society**” than “paying as little tax as possible” ($p = 0.004$).

Conclusion



- Compliance declines (evasion increases) in periods subsequent to a partially-productive audit.
- This decline is moderated by reminder messages of the tax authority's service efforts.
- Both audit productivity and service messaging appears to shift both:
 - Taxpayer focus on fulfilling obligation to society (rather than minimize taxes), and
 - Taxpayer view of the IRS as focused on customer service rather than criminal punishment.
- Limitation: although not identified in this study, boundary conditions certainly exist. Service efforts are unlikely to prevent noncompliance without accompanying detection.

Thank you!





Using the IRS Program Assessment Model Optimizer to Inform Resource Allocation Decisions

IRS/TPC Research Conference
June 18, 2020

*Deandra Reinhart, Clay Swanson, Rafael Dacal, Sarah Shipley, Chris Lee, and Ariel S. Wooten,
Small Business/ Self Employed Research (SB/SE), Internal Revenue Service,
MITRE Team, MITRE*

Where does SB/SE allocate Full Time Equivalents (FTE) across programs?

The SB/SE PAM Optimizer works at the Commissioner-level to inform FTE allocation decisions by:

- Using a data-driven approach that is transparent and repeatable
- Considering impacts of changes to FTE across programs (i.e., Examination and Collection)



The advantage of PAM is that it will allow decision-makers to look at the whole of SB/SE when making staffing recommendations. In this way, they can see simultaneous effects of FTE changes in one program having impacts on other programs.



What is PAM?

- Linear programming model
- Based on a multi-commodity network flow
- Created in R
- Maximizes the collected enforcement tax revenue

Mathematical Formulation

- Objective Function: $F = a_1x_1 + a_2x_2 + \dots + a_mx_m$
- Objective: Maximize or Minimize F
- Constraints:
 1. $b_{11}x_1 + b_{12}x_2 + \dots + b_{1m}x_m (\leq, =, \geq) c_1$
 2. $b_{21}x_1 + b_{22}x_2 + \dots + b_{2m}x_m (\leq, =, \geq) c_2$
 - ...
 - n. $b_{n1}x_1 + b_{n2}x_2 + \dots + b_{nm}x_m (\leq, =, \geq) c_n$
- Bounds: $constant_{i1} \leq x_i \leq constant_{i2}$

The objective of PAM is to maximize collected enforcement revenue generated by working modules of type c at step s

$$\text{maximize } \sum_{sc} r_s^c x_s^c$$

- r is the revenue generated at step s by working modules of type c
- x represents modules of type c processed at step s

- **Total new FTE are less than some total limit N_{max}**

$$\sum_s N_s \leq N_{max}$$

where N_s is the number of new FTE added to step s

- **Executed work is within the capacity of the workforce**

$$\sum_c \frac{x_s^c}{w_s^c} \leq W_s + N_s \quad \forall s$$

where w represents the work rate per FTE, and W_s represents the current FTE

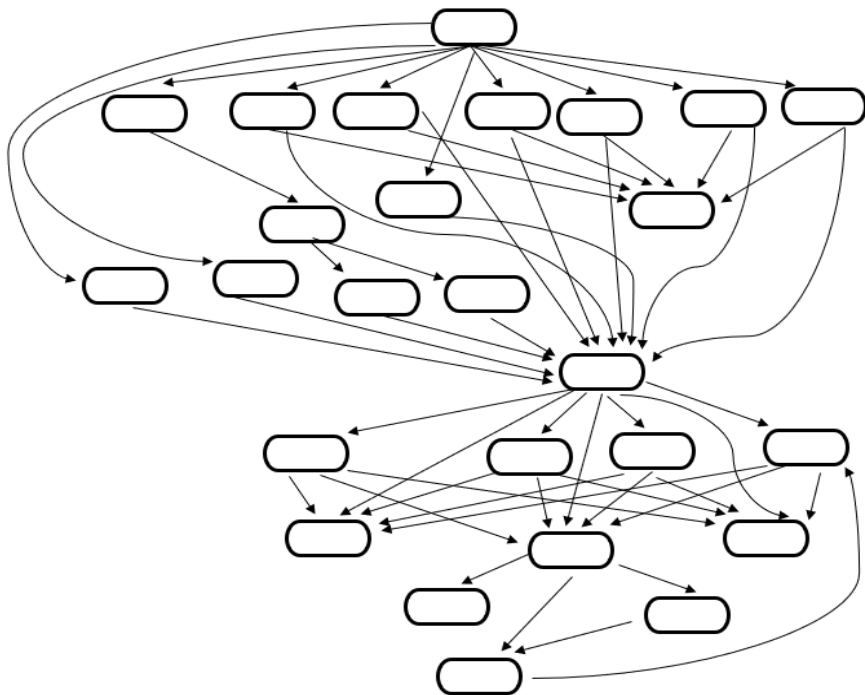
- **Started work must be completed (includes taxpayer correspondence)**

- **FTE minimums and maximums**

$$l_s \leq N_s \leq u_s \quad \forall s$$

- **Work minimums and maximums**

$$l_s^c \leq x_s^c \leq u_s^c \quad \forall s, c$$



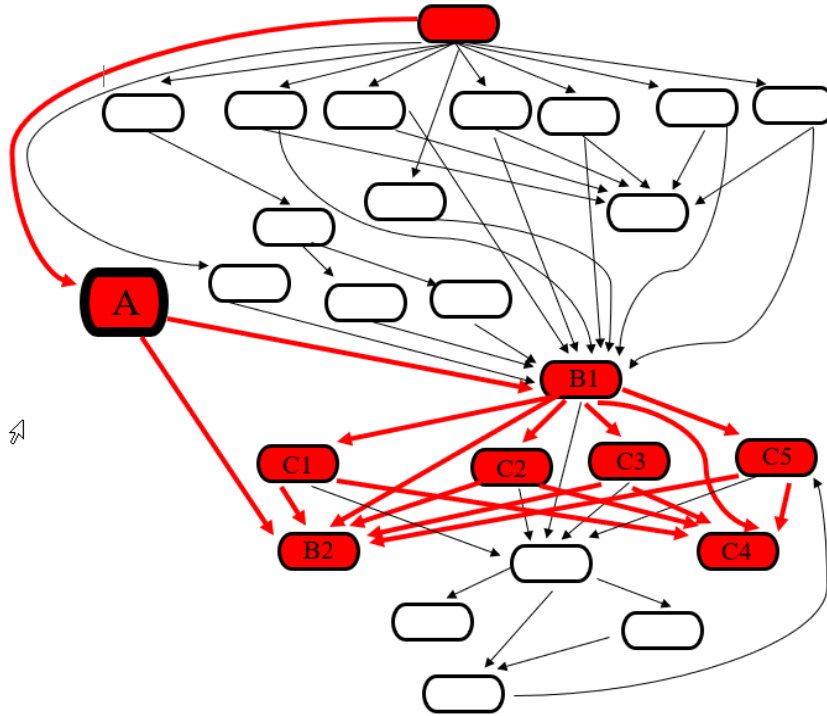
SB/SE programs are interconnected. For example, work identified in Examination creates additional work in Collection, including taxpayer correspondence.

In the network flow diagram, the ovals represent “**steps**”.

The arrows show the flow of “**commodities**” between steps.

- **Step:** a discrete point in the model where work is performed, revenue is earned, and/or work is routed to other steps.
- **Commodity:** a “type” of work that usually represents a particular group of tax modules.

Example: Path of One Type of Commodity



The possible paths for one type of commodity (tax module) are traced in red.

If we add one FTE to step A, we can estimate:

- New work started at step A
- New revenue collected at step A
- Additional work flowing to downstream steps (B1, B2, C1, C2, C3, C4, C5)
- Additional FTEs needed at downstream steps
- Additional revenue collected at downstream steps

PAM compares these results against all other possible FTE allocations for other steps and other commodities.

1

- Solve linear program to optimally allocate new FTEs
- Repeat n times to create n solutions

2

- Cluster n solutions
- Generally 3 to 5 clusters based on analyst judgment
- Solutions in each cluster are averaged

3

- Use cluster allocations to calculate expected revenue
- Repeat m times
- Clusters “compete” against each other

4

- Compare results of m iterations
- Operations judgment
- High yield and robust


More distinction with module types

- Accounts for decreasing marginal value

Workforce evolution

- Expected attrition
- Skill development

Additional optimization measures



10th Annual IRS/TPC Joint Research Conference on Tax Administration June 20, 2020

SPONSORED BY



**Research, Applied Analytics,
and Statistics**

Session 4 Discussion

ARNIE GREENLAND, DISCUSSANT,
UNIVERSITY OF MARYLAND,
AG ANALYTICS, LLC

Thank you for asking me

- It is an honor and a privilege to have been asked to read and think about these three interesting and important papers
- Each helps to answer a core challenge faced by the Service:
 - HOW CAN WE DO MORE WITH LESS?
- The topics include ideas which bring:
 - Better selection of cases for audit
 - New communications approaches during audits to improve compliance in general
 - More optimal allocation of compliance resources

Paper 1: Can Machine Learning Improve Correspondence Audit Case Selection?

Considerations for Algorithm Selection, Validation, and Experimentation

- Selection of cases for audit is one of the oldest and most important problems facing a Revenue Administration
 - And the IRS was a very early user of tools, such as DIF, to provide a mechanism that was accepted by the Examination organization as a useful tool in this process
 - However, as the authors stated, Correspondence audit selection was based on “user developed criteria” or “business rules developed to identify narrow taxpayer populations with suspected noncompliance ...”
- In this paper, the authors investigate using Machine Learning methods to make such selections
 - Focusing on creating ranking and selection tools, with solid theoretical foundations
 - Showing evidence of improvement over “status quo” (existing) methods
 - But also highlighting the important tradeoffs between maximizing revenue collection and lowering the number of No Change Audits

High Level Summary of the Paper

- They described a research process in which
 - An initial research project (Part 1 of the paper) demonstrated potential benefit of using ML tools for (primarily) selection methods
 - Part 2, having the benefit of the findings from Part 1, focused on using a very creative set of ranking tools underlying the process of selecting cases for Correspondence Audit
- My reading of the paper, showed the findings of Part 2 were clearly more encouraging than those of Part 1, but both lend credence to the idea that ML methods can improve the Correspondence Audit selection process

Some highlights of the Pilot

- Part 1 used classic ML techniques for each of two populations
 - Audit Class 1: Taxpayers with Schedule C business expenses
 - Audit Class 2: Taxpayers with Schedule A deductions
- For each Audit Class, the team examined revenue assessed and collected and percent of No Change Audits, comparing the Status Quo (SQ) and Alternative (ML) selection methods
- Findings were interesting but certainly not conclusive, because
 - For Audit Class 1, ML was better than SQ on the No Change dimension, but opposite for the revenue dimensions
 - For Audit Class 2, inconsistencies across years made it difficult to see any clear winners

Some highlights of the Part 2

- Part 2 focused on the Audit Class 1 (Schedule C expenses), but looked at three very interesting methods to rank and order the cases with respect to revenue generated
- The team also used innovative methods to visualize the results
- I found the findings for Part 2 more convincing and interesting
 - Probably the most interesting finding to me was the visualization comparing the Status Quo approach against the ranking method they called “Penalized Regression”:
 - Looking at both Lift chart which compared cumulative counts (in order of the new ranking versus cumulative revenue) and the “No Change Progression Plot (counts in rank order versus No Change Rate), the **Penalized Regression was better than Status Quo approach in both cases.**
- This finding should drive some additional research and possibly a new selection method for Correspondence audits

Overall

- I really enjoyed this paper
- This is very important work, well written and presented.
- I hope that the Service will continue to develop and test tools of this nature

Paper 2: Audit productivity, taxpayer service, and compliance: Can a service mindset overcome a dwindling enforcement budget?

- The authors focus on a core hypothesis (and 2 specific components of that hypothesis):
 - More positive messaging about the service aspects of the IRS during an audit can improve downstream compliance
- They also examined the issue of whether an audit that is limited in scope or effectiveness (not 100% productive) might result in increased noncompliant behavior by that audited taxpayer
 - But also whether more positive service messages could lessen that impact
- To test these hypotheses the authors created a research design using ideas of experimental economics as a guide

Summary

- The two core hypotheses were:
 - H1: Taxpayers who experience partially-productive audits are less likely to increase compliance in subsequent years than those who experience fully-productive audits.
 - H2: Taxpayers who experience partially-productive audits are more likely to increase compliance in subsequent years when they also experience increased service efforts.
- To test these hypotheses, they designed a testing protocol and recruited participants using Turk Prime (a Google affiliated service) that allows potential respondents to be vetted, and for minimal remuneration perform tasks required by the designers.
- Through that resource they obtained 271 usable cases, across a multi-dimension design allowing for different levels of audit productivity, different messaging treatments, and through a series of compliance activities, sequenced in time, to be performed by the respondents

Continued summary

- The key metric was the change in compliance measured by the amount of their “experimental” income truthfully reported as income
- The authors used standard multi-variate statistical analysis to evaluate whether the changes were significant and in the direction that proved out their hypotheses
 - In addition, they employed a very interesting approach (Contrast Testing), developed in the Accounting literature, to focus more directly on the interactions of the respondent behaviors in support of their hypotheses in the experiment

Some comments

- The key findings were that both hypotheses were supported in the analysis of their experiment with the following message:
 - Positive messaging about the service aspects of the IRS during audits can improve compliance in general, but also mitigate the negative impulses of continued non-compliance for taxpayers who experience a less than 100% productive audit.
- This is a **very important message**, consistent with the title of this Session, and for managers at IRS to consider when developing improved audit protocols
- The methods used were both interesting and innovative
 - I found myself digging into the paper by Guggenmos, *et. al.* to learn more about just exactly how Contrast Testing works

Some Concerns

- I was not fully comfortable with the use of paid respondents, obtained through an online source
 - I realize that the authors did provide screening protocols, and evaluated demographic characters (mentioning, for example, the higher education of participants)
 - However, the demographics also showed many other disparities from the larger population, for example in age and income
 - I would have liked to have gained a better understanding of the historical tax compliance patterns of the specific sample they had
- Also, I found the sample size just a bit low, at 271 usable cases, spread cross 4 strata, leaving some strata having as low as 46 cases.
 - To me, this suggests the need for confirmatory research to bolster credibility of the findings

Overall

- Another excellent paper
- One that focuses on the core theme of this session of DOING MORE WITH LESS
- Containing ideas that cry out for attention by IRS managers, and designers of examination protocols

Paper 3: Using the Internal Revenue Service Program Assessment Model Optimizer to Inform Resource Allocation Decisions

- The paper deals with a very important problem, and one which I also once worked on myself
 - Allocating resources within the IRS Compliance function in a more optimal way, or more correctly putting IRS compliance resources where they will best meet the Service's strategic goals
 - We have already learned in other papers in the session that increase revenue and the tradeoff with NO CHANGE audits is core to that thinking
- The model uses the powerful tool of Mathematical Optimization, but also other analytics tools such as Monte Carlo simulation and cluster analysis as part of the entire package
- I have many questions, mostly about the actual operation of the model, but also on the form, experience in its use, and validation of the findings

There are many positives in the paper

- The paper describes a very robust optimization model
 - Taking into consideration the multi-stage characteristic in which work done at earlier stages feeds the downstream activities at later stages
 - Devising a simulation approach that models the variation in workload completion rates, complexity of work, and other uncertainties in the work
- My personal assessment is that this model, if implemented, will increase efficiency and improve key metrics such as revenue collection, both bringing much value to the Service

Issues and questions

- First a CAVIAT:
 - I expect that many of the issues/questions I had will be addressed in the presentation, in which case, I will not focus on them in my verbal comments
- The first issue focuses on the basic unit of analysis being revenue/FTE. While this makes sense notionally and may be more convenient, FTEs are not fungible across program categories (e.g., an RA FTE working a complex field audit costs a lot more than a correspondence FTE working simpler issues). I believe the model should be restructured using, for example, revenue/budget \$ allocated.
- The model shows many variables, which it seemed to me were naturally INTEGER variables, these include:
 - Required and discretionary number of modules worked
 - Number of FTE's added
 - While approximating them with possibly fractional values is often done, the rationale for those assumptions, and sometimes a sensitivity analysis to support that decision, is typically provided

Continuing with issues and questions

- I typically also like to understand the size and complexity of the model which would include:
 - Number of variables, number of constraints, solution approaches
 - What was the solution speed, and was there the need to use advanced methods such as decomposition, column generation, or the creation of cutting planes?
- Another area that I find of interest is the process of creating the model:
 - How were the goals of the model determined?
 - Are there multiple goals, some that might be tradeoffs (e.g., the Revenue/NoChangeAudit issue)?
 - Did multiple organizations within the IRS participate and were they part of creating of the constraint set?
 - This relates to the issued raised earlier about allocation of dollars rather than FTEs.
 - Did new constraints emerge during model development?
- Finally, what has been the process of socialization/acceptance of the model?

Overall

- This was a very interesting and important paper
- I am sure that there are good answers to the questions I posed, and I hope that this work goes forward to implementation and acceptance by the key players at the Service



Session 4. Doing More With Less

Moderator:

Tom Hertz

IRS: RAAS

Can Machine Learning Improve Correspondence Audit Case Selection? Considerations for Algorithm Selection, Validation, and Experimentation

Ben Howard

MITRE Corporation

Audit Productivity, Taxpayer Service, and Compliance: Can a Service Mindset Overcome a Dwindling Enforcement Budget?

Mary Marshall

Louisiana Tech Univ.

Using the Internal Revenue Service Program Assessment Model Optimizer to Inform Resource Allocation Decisions

Clay Swanson

IRS: SB/SE

Discussant:

Arnie Greenland

A. G. Analytics, LLC