Will AI Accelerate the Digital Divide?

Evidence from a Field Experiment on How Managers use Explainable AI

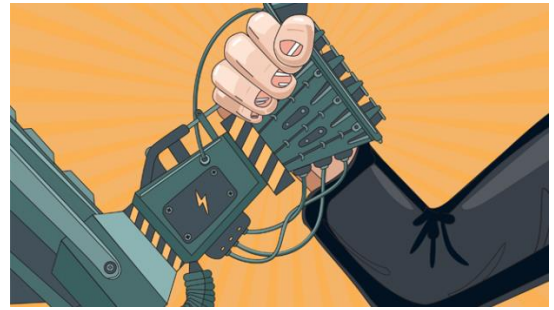
Selina Carter
Carnegie Mellon University
Department of Statistics

Jonathan Hersh
Argyros School of Business
Chapman University

Introduction

- **Machine learning** predictions have been used to support decisions in:
  - Banking and finance (e.g. credit risk evaluation, Baesens, B. and Setiono, 2003)
  - Transportation (e.g. Train delays, Oneto, et. al 2018)
  - Medicine (e.g. ICU admission Yoon, et al., 2016).
  - Education (e.g. at-risk dropouts Lakkaraju et al., 2015)

- However these nearly every example supports decisions that are individually of relatively low-stakes, and often designed for experts.
- **Does role within a firm or ML familiarity affect worker trust in machine learning predictions for high-stakes decisions?**
- **Does Explainable AI affect algorithmic trust?**
- **Who benefits within a firm from AI and Explainable AI?**



Introduction

- We use machine learning predict delays in loan execution for a large multilateral bank
  - We embed these into a dashboard and survey 685 employees
- **Randomized treatment**: Some workers receive an Explainable AI module to “explain” AI predictions
- **Results**: Mixed evidence explainable AI affects AI trust overall
  - Senior managers and self reported ML novices least likely to trust AI
- However, AI reluctant groups respond strongly to explainable AI
  - Suggest convergence of AI adeptness

Outline

1. Introduction
2. Background and Literature Review
3. Managerial Context
4. Field Experiment on AI Trust
5. Experiment Results
6. Conclusion

Why Do We Care About Explainable AI?

1. Consequential decisions may need human validation
2. Explaining the model may help us build better models
3. Explainable models might engender more trust

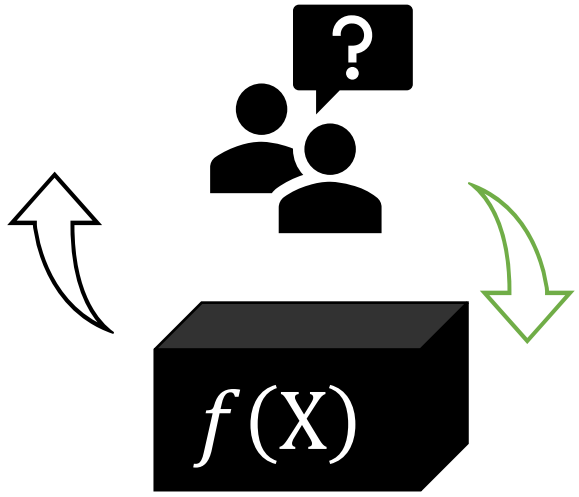

Opinion

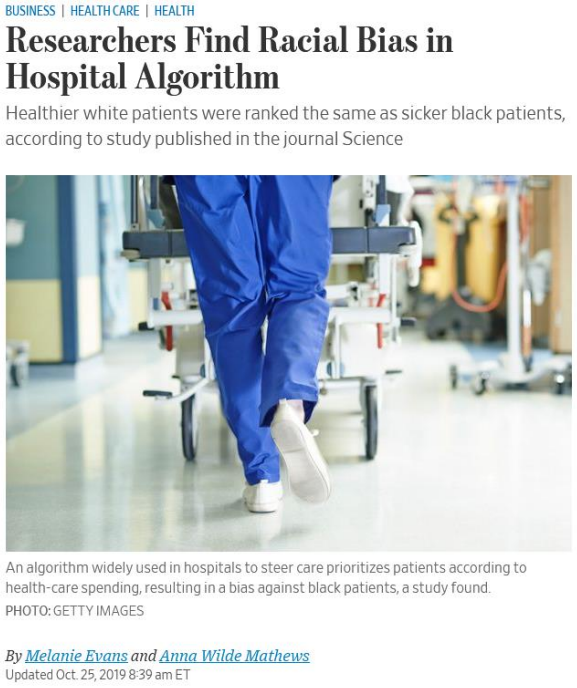
OP-ED CONTRIBUTOR

When a Computer Program Keeps You in Jail

By Rebecca Wexler

June 13, 2017




BUSINESS | HEALTH CARE | HEALTH

Researchers Find Racial Bias in Hospital Algorithm

Healthier white patients were ranked the same as sicker black patients, according to study published in the journal Science

An algorithm widely used in hospitals to steer care prioritizes patients according to health-care spending, resulting in a bias against black patients, a study found.

PHOTO: GETTY IMAGES

By Melanie Evans and Anna Wilde Mathews

Updated Oct. 25, 2019 8:39 am ET

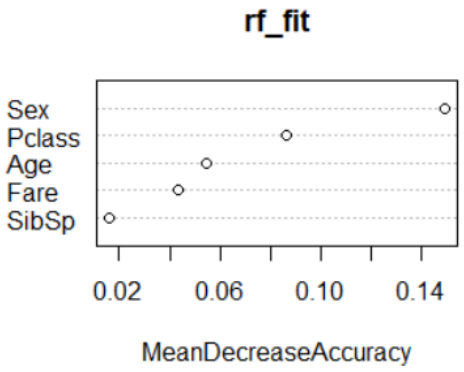
What is Explainable AI?

- Explainable AI is artificial intelligence where the computer's decisions can be understood by a human
- Often start with a black box model and ask computer to "explain"

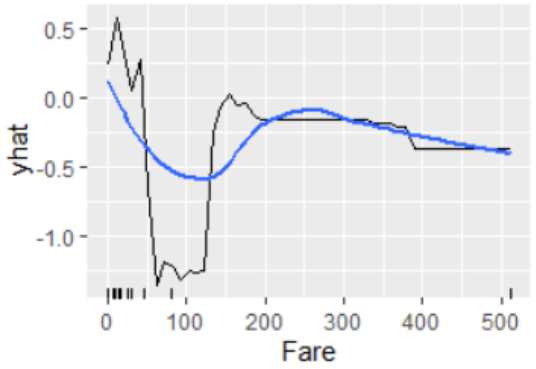
Maybe our models are too complicated? Angelino, Larus-Stone, Alabi, Seltzer, and Rudin. **Learning Certifiably Optimal Rule Lists for Categorical Data**. JMLR, 2018. https://youtu.be/sl78EgrT4TY

Global feature importance

(Which variables drive the results)



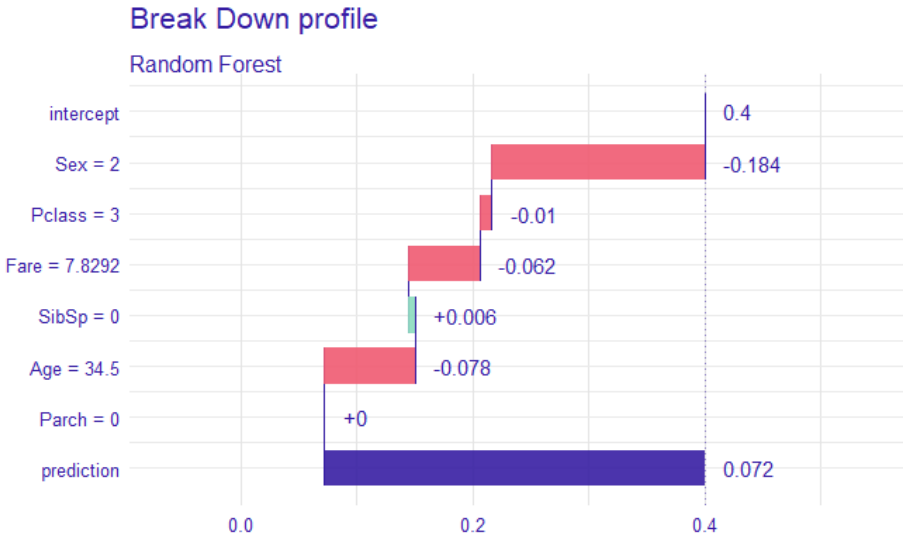
Variable Importance



Partial dependence

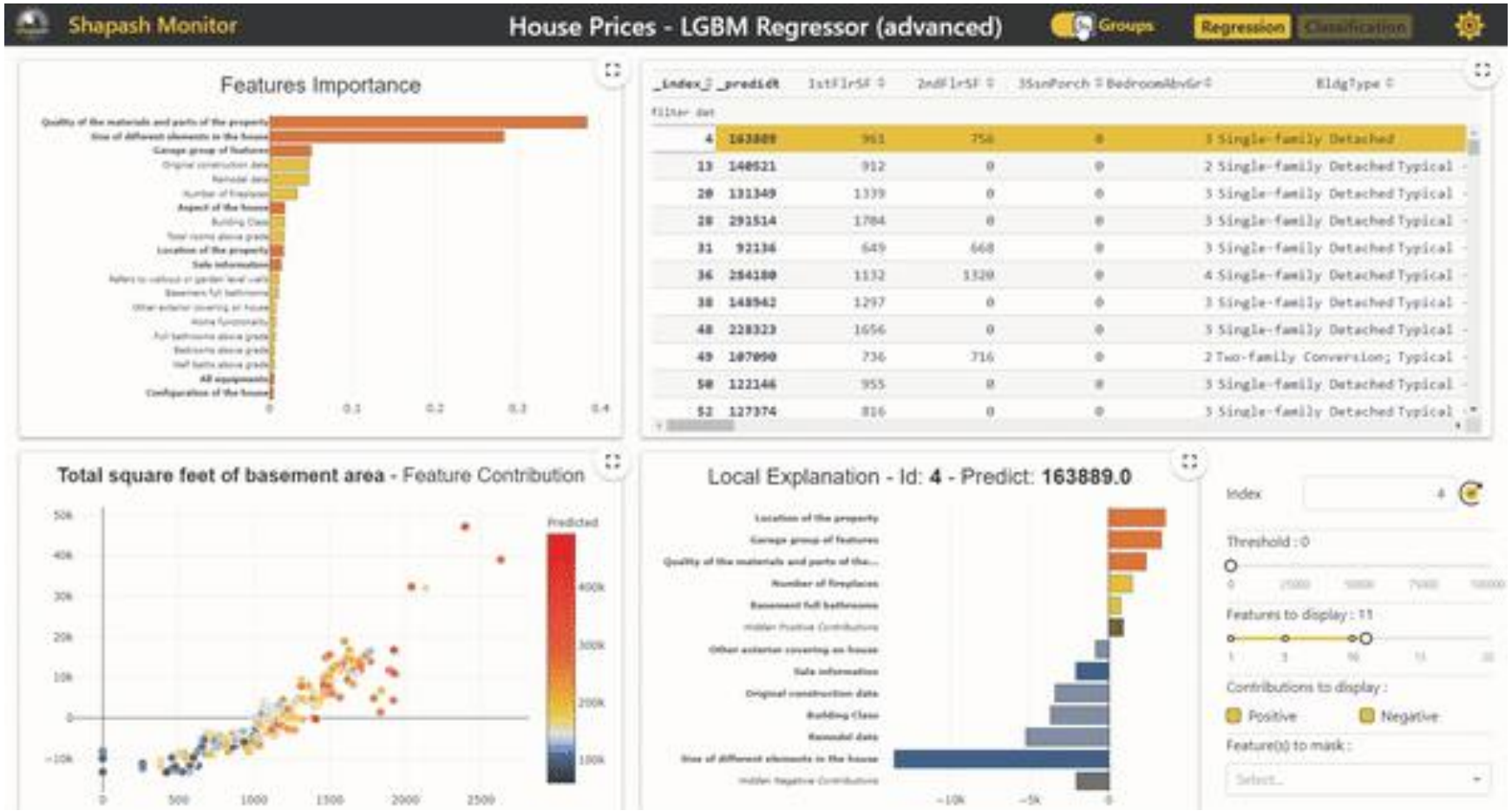
Local feature importance

(Explain prediction for a given obs)



Shapley plots

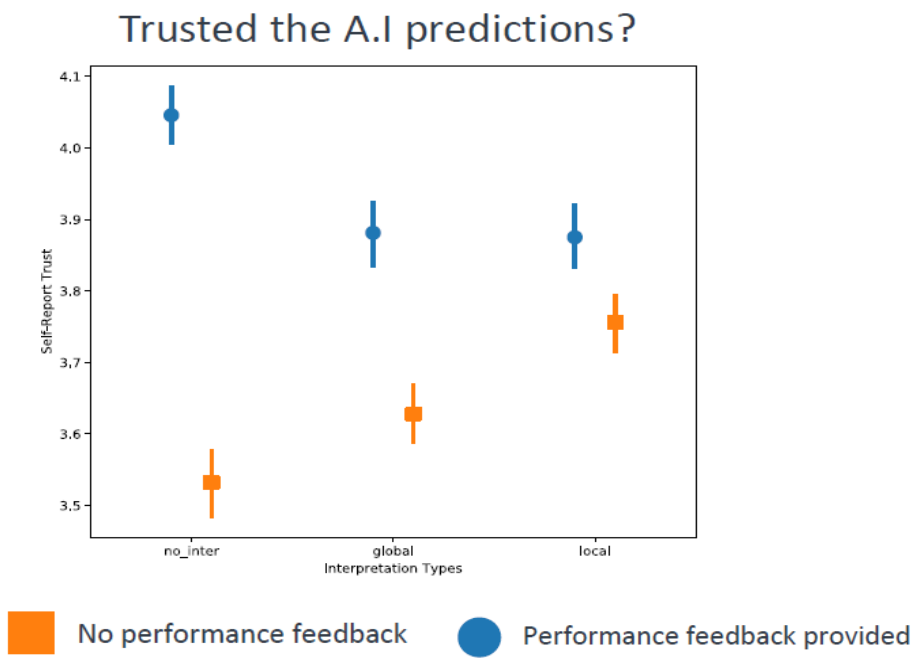
AI “Dashboards” Are Often Employed to Explain Predictions



https://github.com/MAIF/shapash

Literature 1: What Drives Algorithmic Trust?

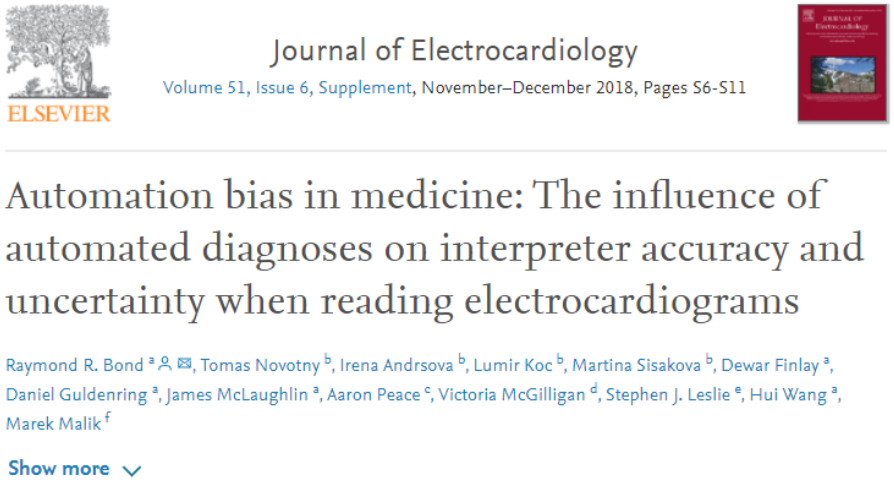
- **Ahn, Almaatouq, Hosanagar (2020):** Possible drivers of algorithmic trust:
  1. **Transparency:** how does the algorithm work?
  2. **Interpretability**: Can the model explain its own decisions in ways that humans can understand?
  3. **Performance**: Is performance sufficient and communicated?
  4. **Control:** Can the user modify the algorithm
- Lab experiment to test trust speed dating algorithm
  - Global and local importance increase trust
  - Performance feedback also increase trust



Literature 1: Algorithmic Trust Anomalies

Low trust persists for fully-automated decision systems

- E.g. vehicle piloting systems (Dikmen and Burns, 2017) Robotic surgery (Sullins, 2014)
- Trust may be low even when machine predictions dominate human predictions e.g. jail decision Kleinberg et al., 2018).

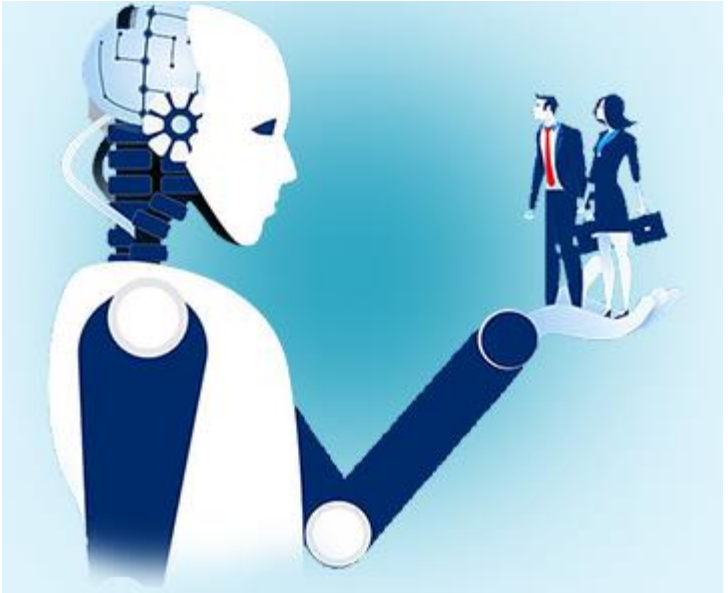


Automation Bias can anchor decisions to machine suggestion

- Evidence of automation bias for medical personnel viewing electrocardiogram-based automated diagnosis (Bond et al., 2018)

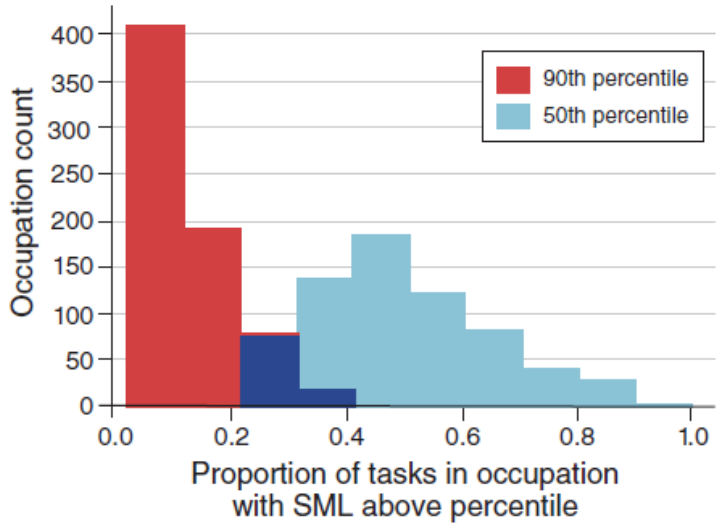
Literature 2: AI Changes How Firms Make Decisions

- Dixon, Hong and WU (2021) find that robots increase firm employment but decrease total managers. This centralizes decision-making authority (contrary to earlier IT innovations)
- Van den Broek, Sergeeva, and Huysman (2021) show that AI system developers rely on domain expertise in an iterative fashion.
- Fügener et al. (2021) finds that humans interacting with AI often behave like “Borgs”. Variance among human decisions decreases and “wisdom of crowds” is lost.
- Tong, Jia, Luo and Fang (2020) find that AI feedback may increase worker productivity, but harms worker productivity when it is revealed an AI generated the feedback



Literature 3: Technology Re-Skilling Gap

- Brynjolfsson, Mitchell, and Rock (2018) look at occupational tasks and which are "suitable for machine learning" replacement
- Most jobs have some, but not all, tasks that require ML
- Instead of job replacement: think job redesign
- How will workers be retrained for those redesigned jobs?



Tom Relihan | Jun 26, 2018

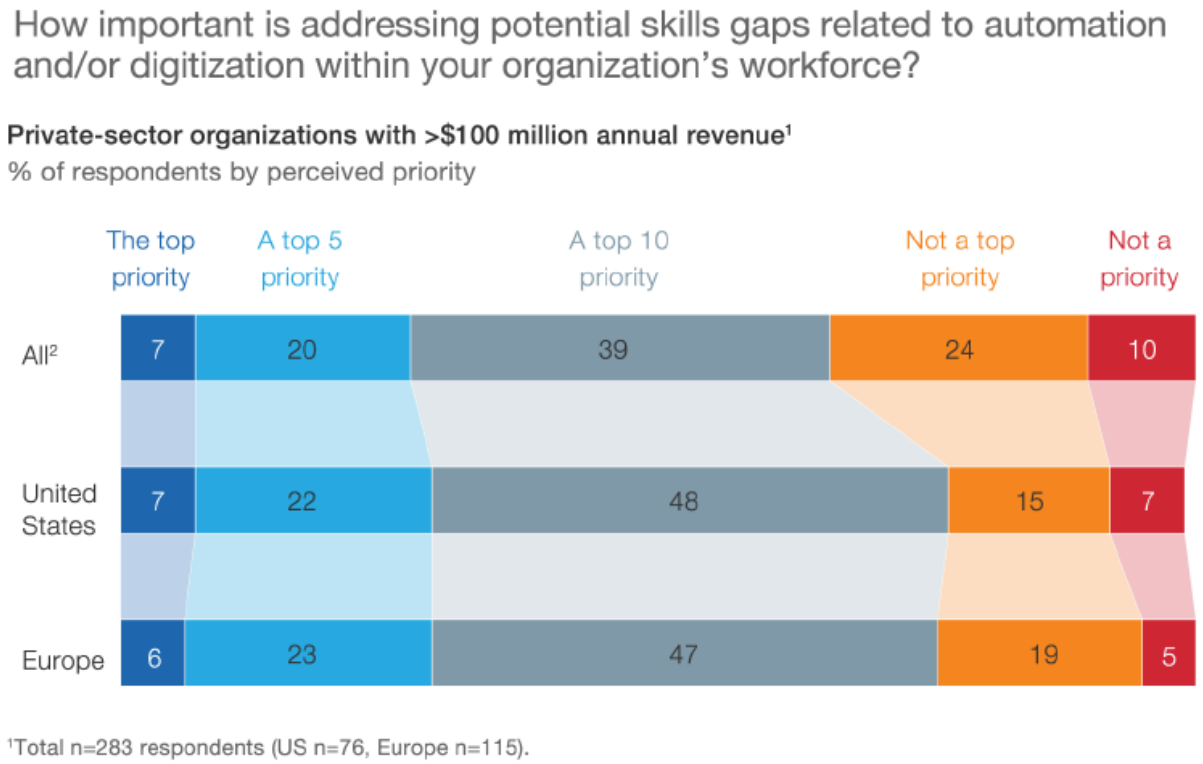
Why It Matters

It's time to shift the conversation around AI and machine learning from threats of job replacement to opportunities for job redesign.

Share

Literature 3: Technology Re-Skilling Gap

- Brynjolfsson (2019) recommends “reinventing education” around STEM, technology, art and design.
- McKinsey Global Institute projects 375 million workers may need to retrain as AI changes firm demands for workers.



[1]Total n=283 respondents (US n=76, Europe n=115).

Literature 3: Technology Re-Skilling Gap

- Organizations aren't particularly focused on retraining for AI
- Will this happen organically? (Think email, Excel)

Empirical and Experimental Setting

- We partnered with **a major development bank** to build a machine learning decision support tool to predict delays in execution of *sovereign guaranteed investment loans*.
  - Think: large infrastructure loans, avg size $67m USD.
  - Delays very costly: Only 22% loans disbursed on time, 24% of supervision cost from delays

Bank Balance Sheet Loan	
Average loan size	$67 million
Loan approvals	≈ 90 new loans per year
Loans in execution	≈ 500 loans at any time point
Loans with delay meeting project deadline	**78%**
Average delay	**14 months**

Example Project: Salto Grande Hydroelectric Dam Modernization

RG-L1124 : Modernization of the Salto Grande Binational Hydropower Complex

Project Status: Implementation

The overall objective is to help ensure the availability of the Salto Grande Hydropower Complex(SGHC), enhancing the reliability and efficiency of the interconnection between Argentina and Uruguay. The specific objective is to assist in extending the useful life of the SGHC by modernizing its infrastructure and equipment

PROJECT DETAIL

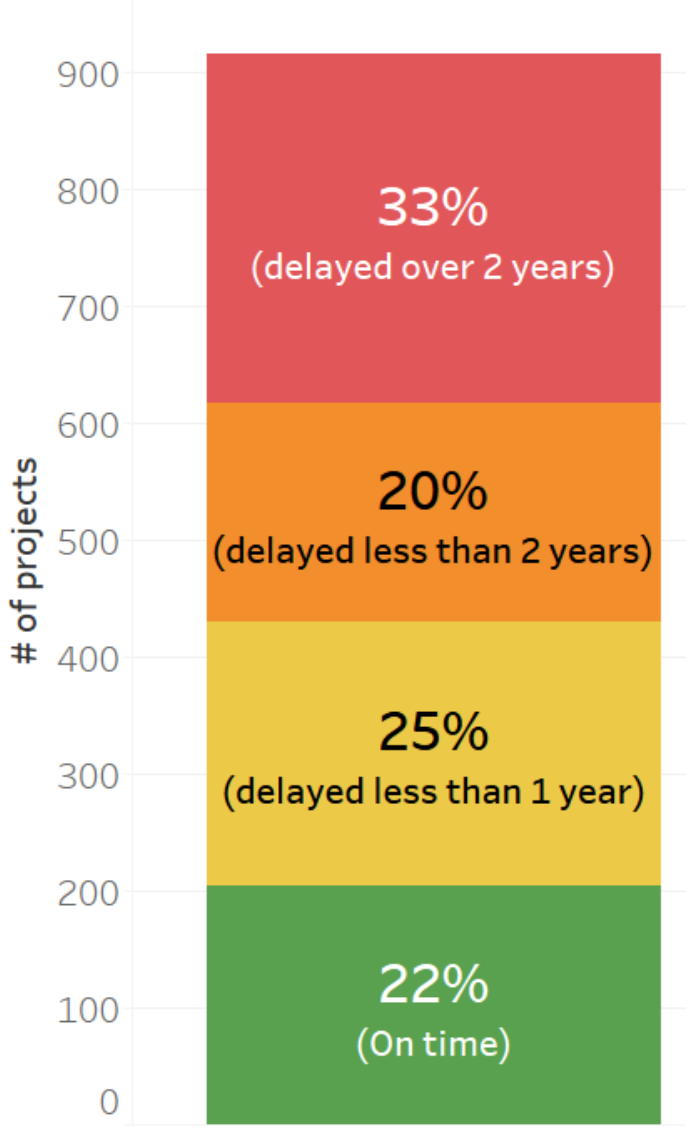
PROJECT NUMBER	RG-L1124
APPROVAL DATE	November 28, 2018
PROJECT COUNTRY	Regional
PROJECT SECTOR	ENERGY
PROJECT SUBSECTOR	ENERGY INTEGRATION
PROJECT TYPE	Loan Operation
ENVIRONMENTAL AND SOCIAL IMPACT CATEGORY	B
PROJECT STATUS	Implementation
OPERATION NUMBER	4694/OC-RG ⊕
OPERATION NUMBER	4695/OC-RG ⊕

PROJECT INFORMATION

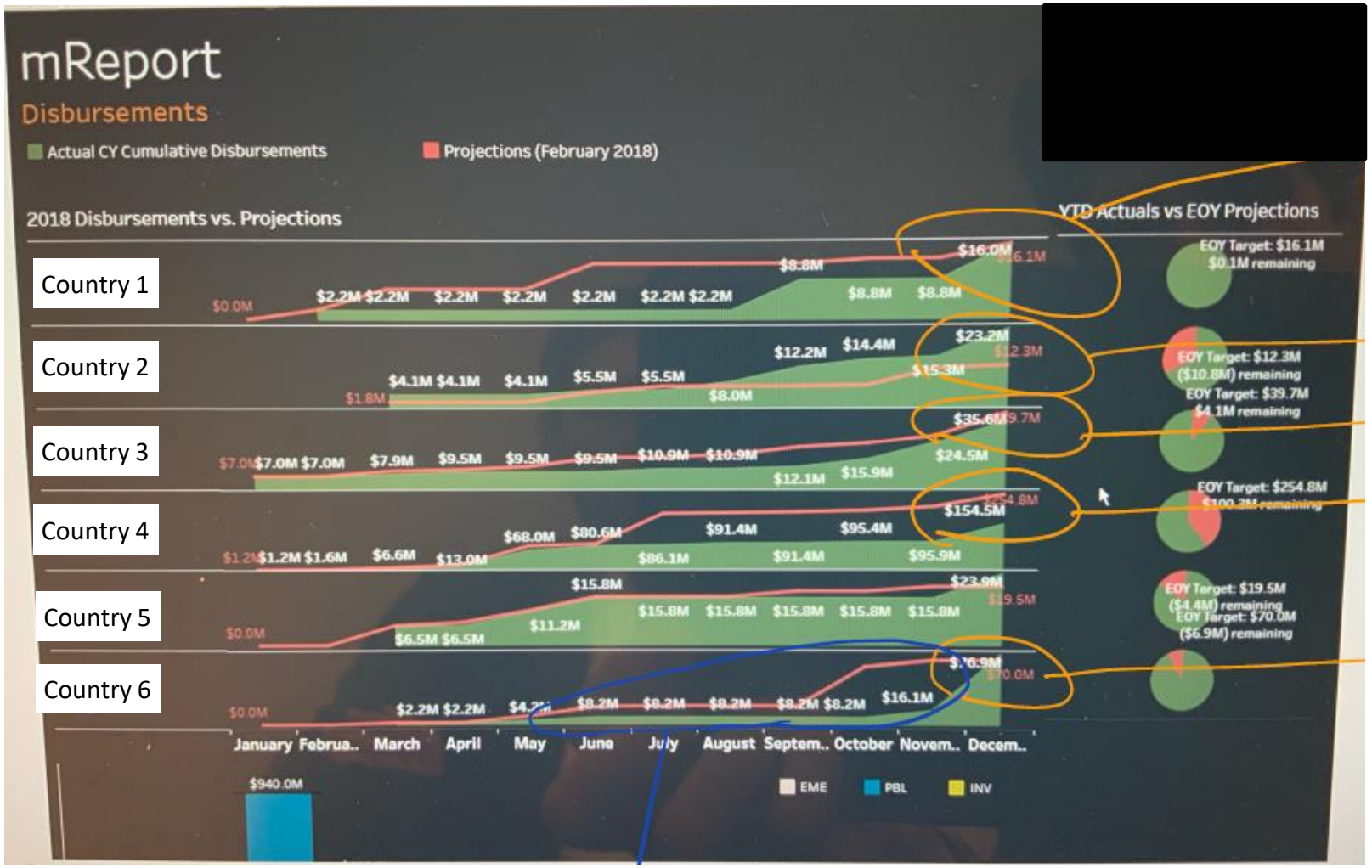
TOTAL COST	USD 80,000,000
COUNTRY COUNTERPART FINANCING	USD 0
AMOUNT	USD 80,000,000

Why is Predicting Loan Delays Useful?

- **Only 22% of loans executed on schedule**
  - Average extension = 14 months
  - **33%** extension > 24 months.
- **What does this cost?**
  - **$50 million** on the supervision of extended loans between 2010 and 2017.[1]
  - = **24%** of the total direct costs for supervision[1]
  - An extended project costs **59%** more on average than on-time projects



Current System: Yearly Report on Projected Disbursements



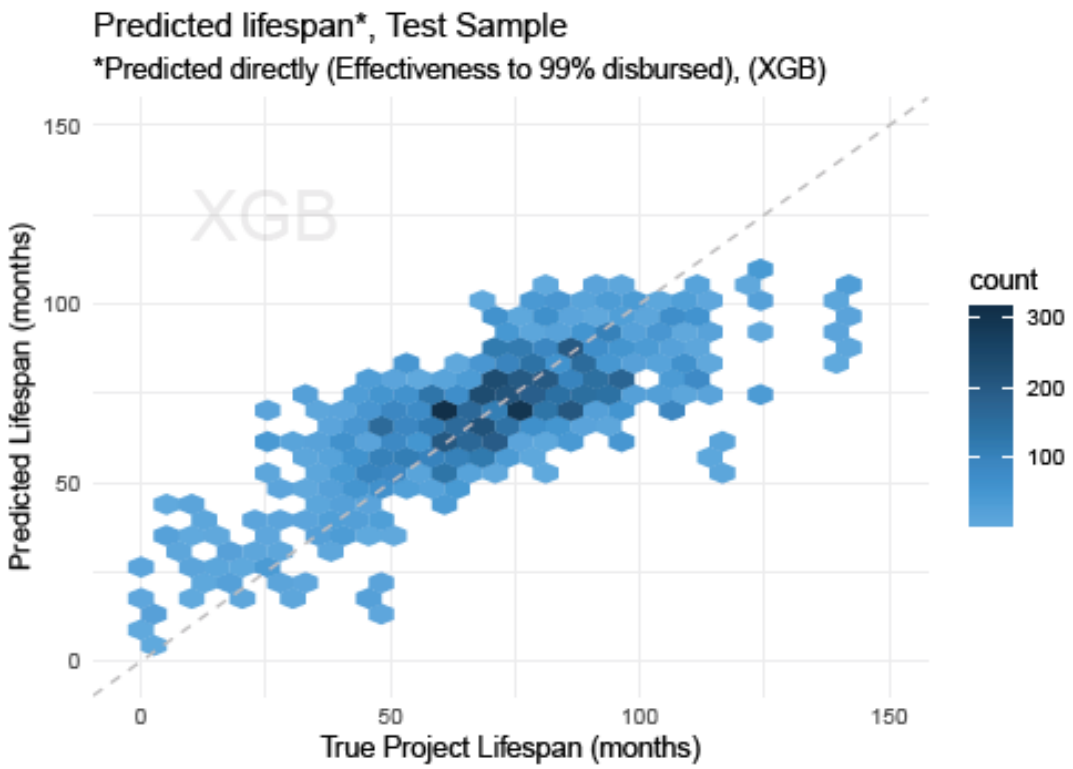
Machine Learning Model

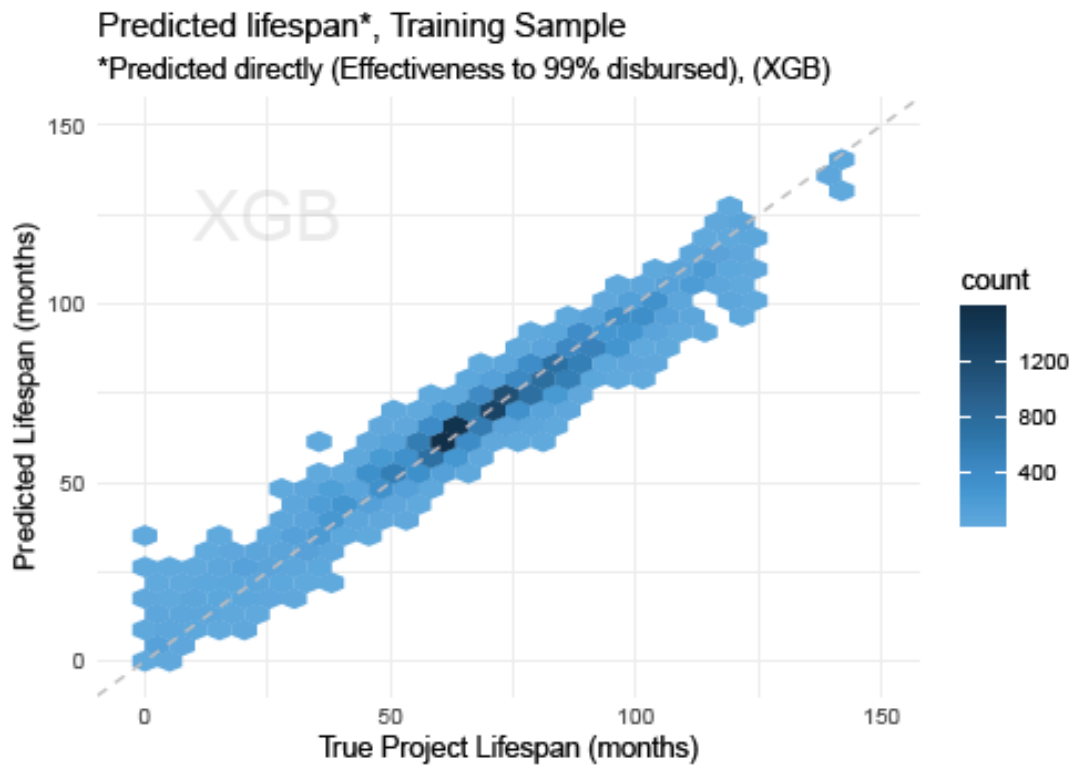
- **Training/Testing Data**: ≈ 1,500 loans approved from 2000-2019
  - 85% Train / 15% Test at project level
- **Outcome variable**: months a project is delayed
  - y_i = delay (in months) between the *original* deadline for disbursing all project funds and the *actual* date at which the project disburses 95% or 99% of total funds
- **Predictor variables**:
  - X_i. = project-level information (100+ variables)
    - **Non-temporal**: recipient country, economic sector, size ($), type of local partner (central government, state, municipality, other), etc.
    - **Temporal**: Cumulative disbursement history,, number of changes in the project team leader, standardized score of project results over time, etc.

- **Model:** Extreme Gradient Boosted Trees

Mean Absolute Error (MAE)	R^2	Testing/Training
11.16	0.536	Test

Model accuracy and diagnostics





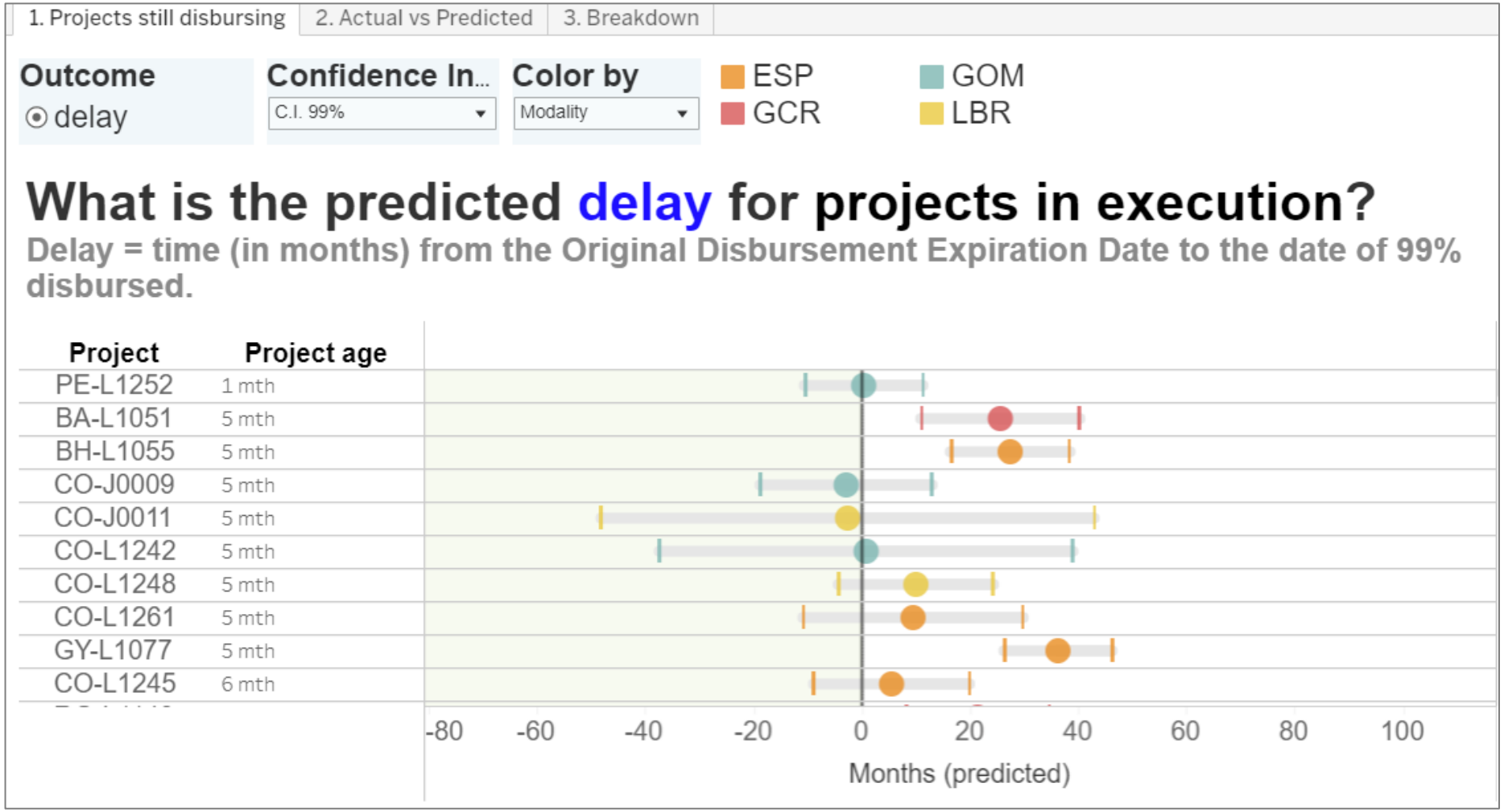
Field Experiment Setup

- Every February, team members must update disbursement forecasts for every loan

- **Sample frame:** 685 team members
- **Timeline:** Survey active May-June, 2021
- **Responses**: 617 (90%), 490 with active projects (sample)
- **"Control" group:** Dashboard with delay predictions ($N_c = 263$)
- **"Treatment" group:** Dashboard w/ predictions, model accuracy and explainable AI ($N_T = 227$)

- **Employee Roles**:
  - Chief of Operations (overseeing loan portfolio for country)
  - Team Leader (manages a particular loan)
  - Analyst (various technical tasks)
  - Fiduciary/Procurement (due diligence)
  - Other (administrative support)

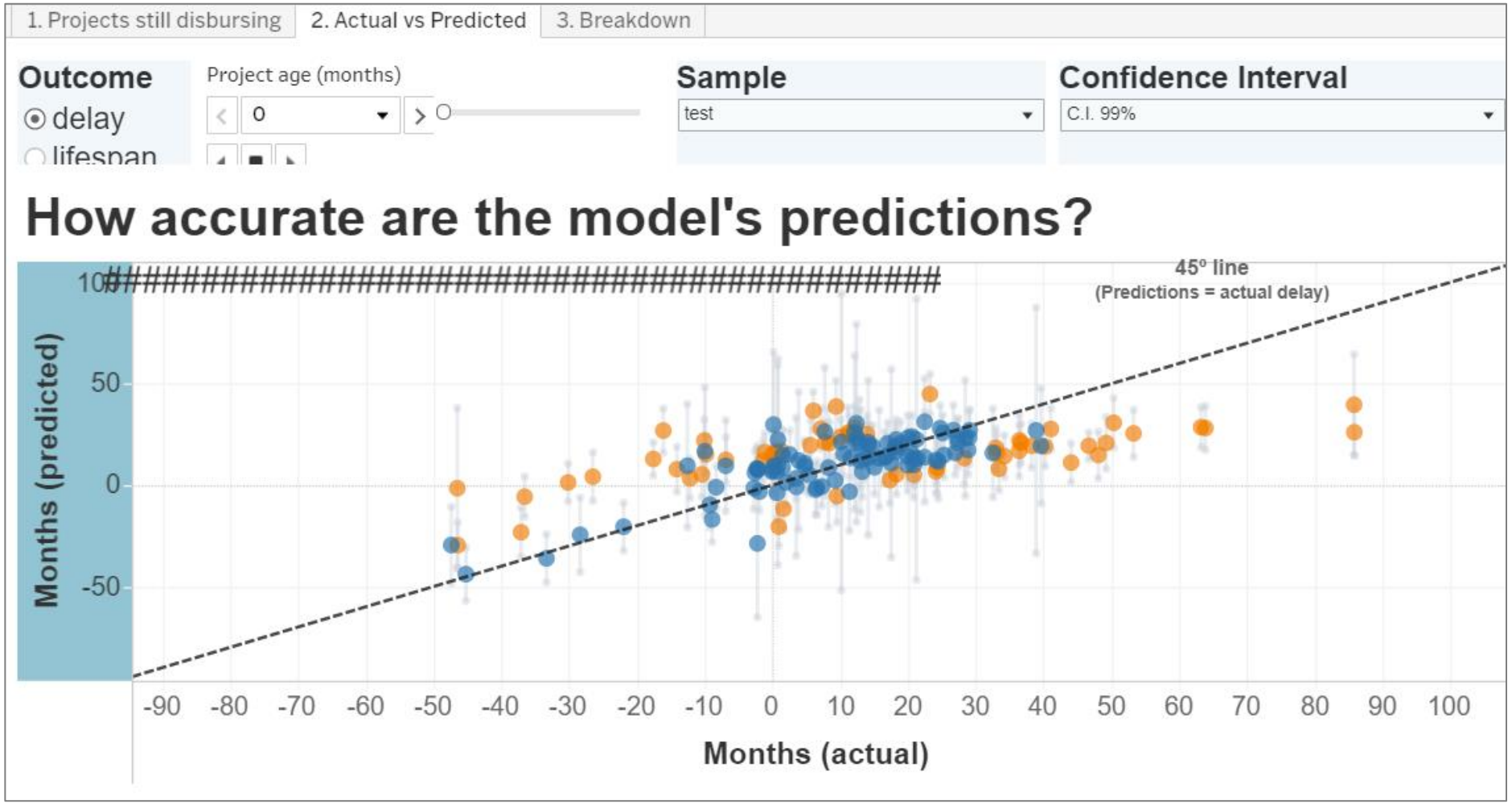
“Control” Group Dashboard



Control group:
$(N_{control} = 263)$

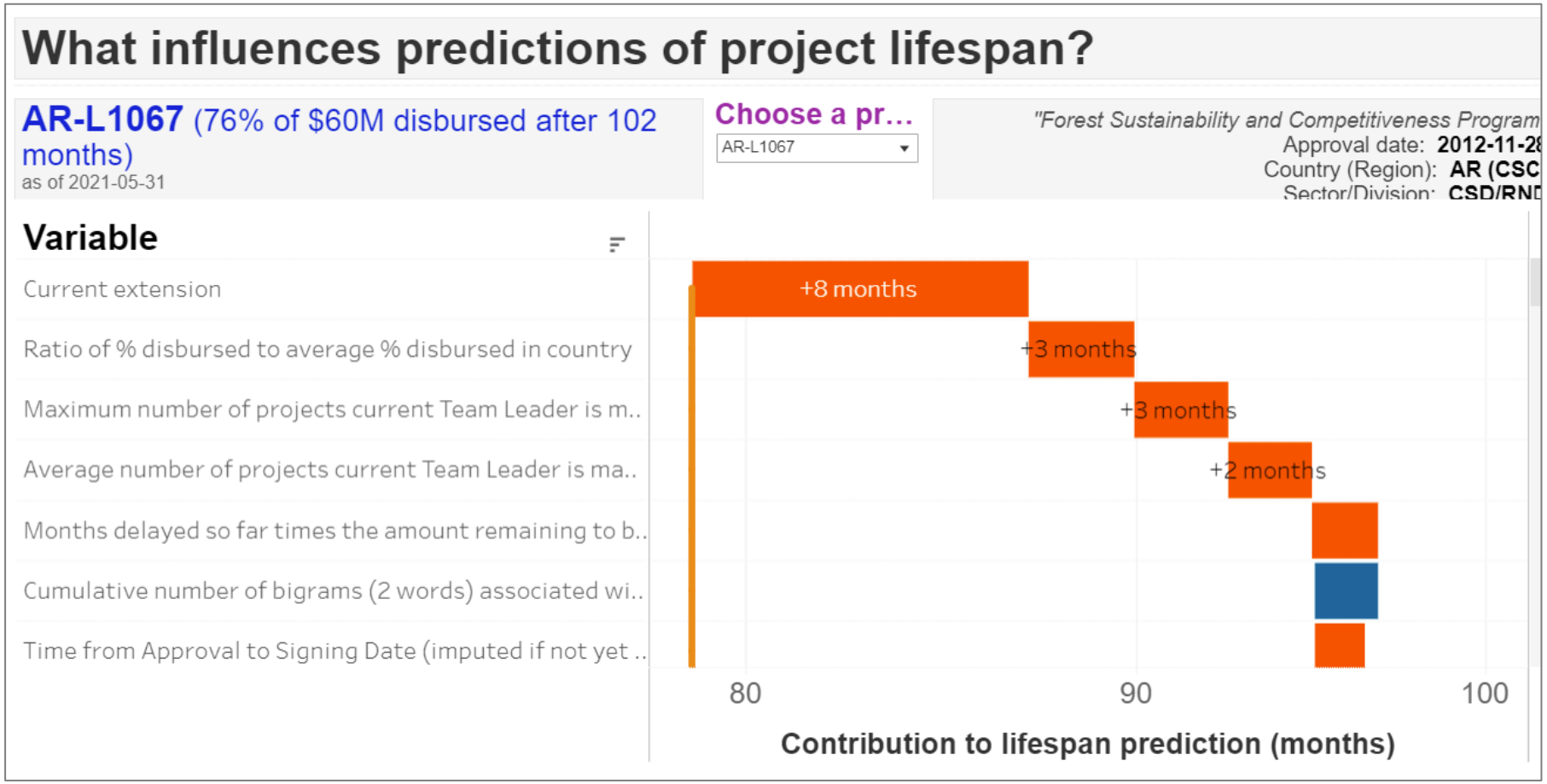
Dashboard with prediction and confidence intervals (via infinite jackknife Wager et al., 2014)

Explainable AI / Model Performance Treatment Group Dashboard



Model performance Panel

Explainable AI / Model Performance Treatment Group Dashboard



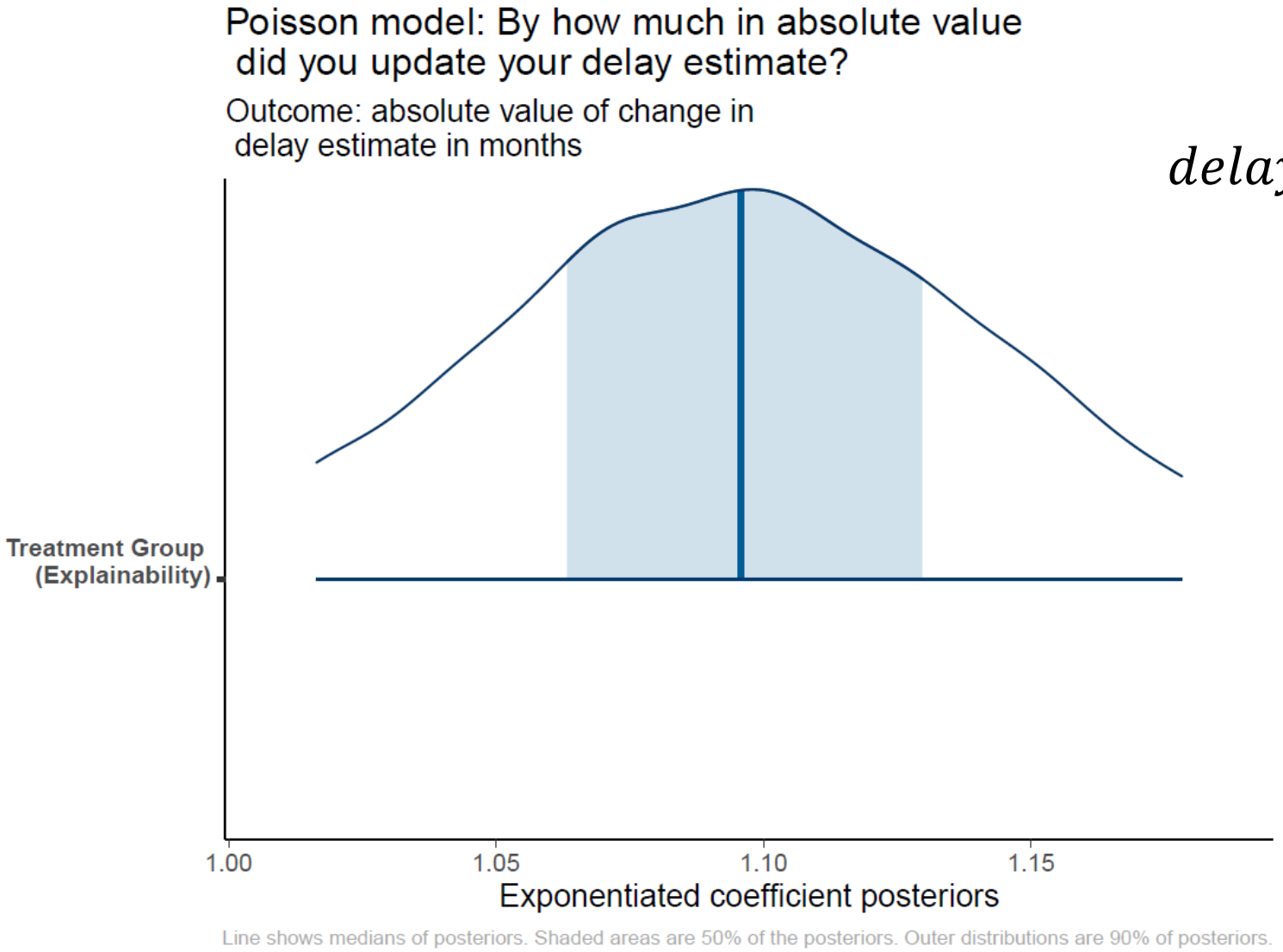
Shapley local feature importance about why a given observation was given a certain prediction

Results Analysis

- **Four outcome variables:**
  - Did user change their delay estimate before/after viewing tool (=1 if updated)
  - Absolute value of change in delay estimate update
  - How useful is the ML tool? (Ordered rank 1-5)
  - How well did you understand the ML tool? (Ordered rank 1-5)

- **Five Moderators** (Pre-registered AsPredicted.org #69245)**:**
  - Explainable model/performance dashboard
  - Work Location
  - Role (Team Leader, Analyst, Fiduciary/Procurement, Chief of Operations)
  - Loan Amount
  - Machine Learning Familiarity (1 least, 5 most)

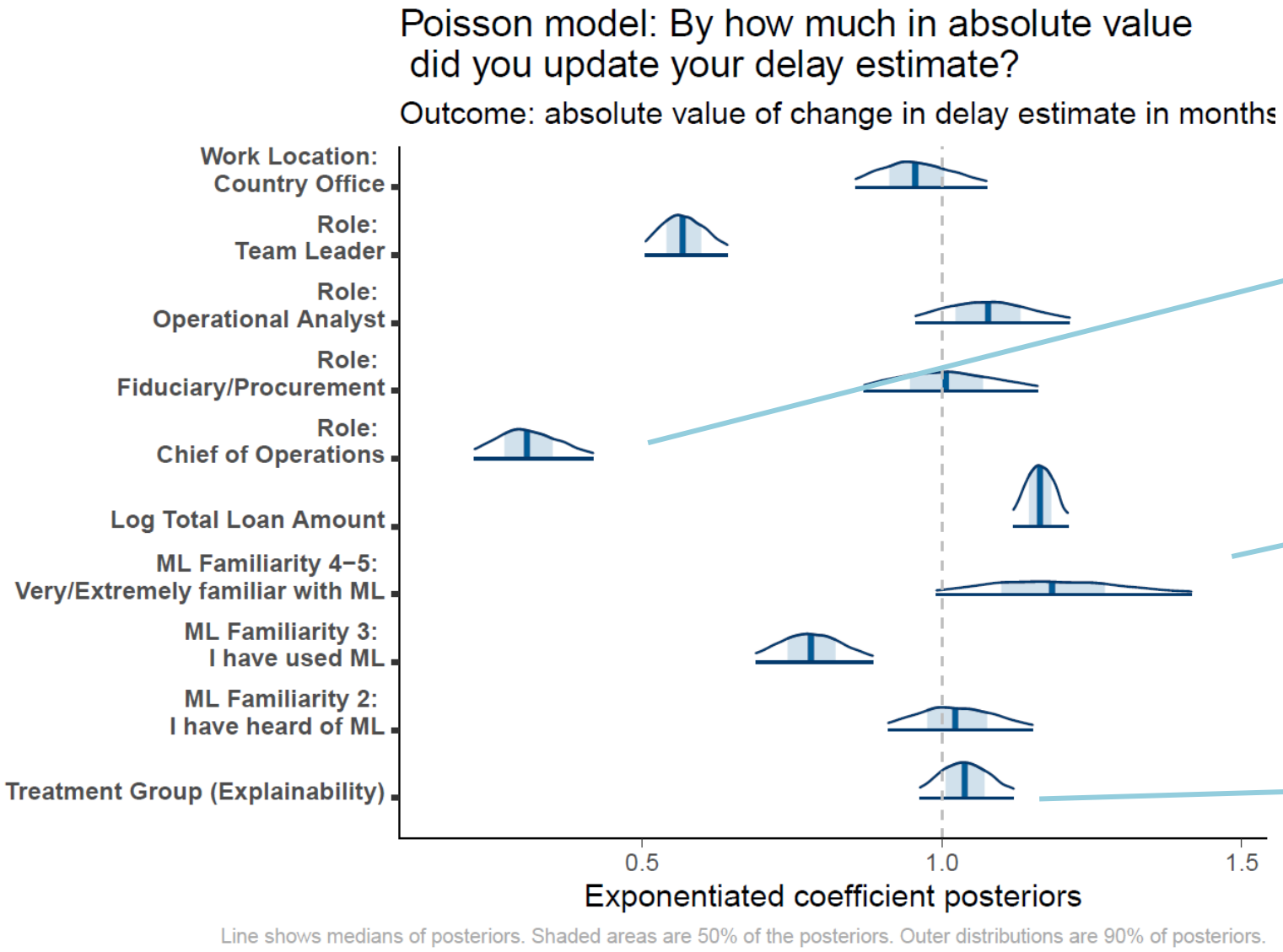
- **Bayesian Ordered Logistic, Poisson, or Logistic Depending on Outcome**



$$delay\_update_i = \exp(x_i' * \beta_{Treatment})$$

- Median: $\exp(\beta_{Treatment}) = 1.1$

  Average delay update = 2.2 months -> 4.5% impact of explainability treatment



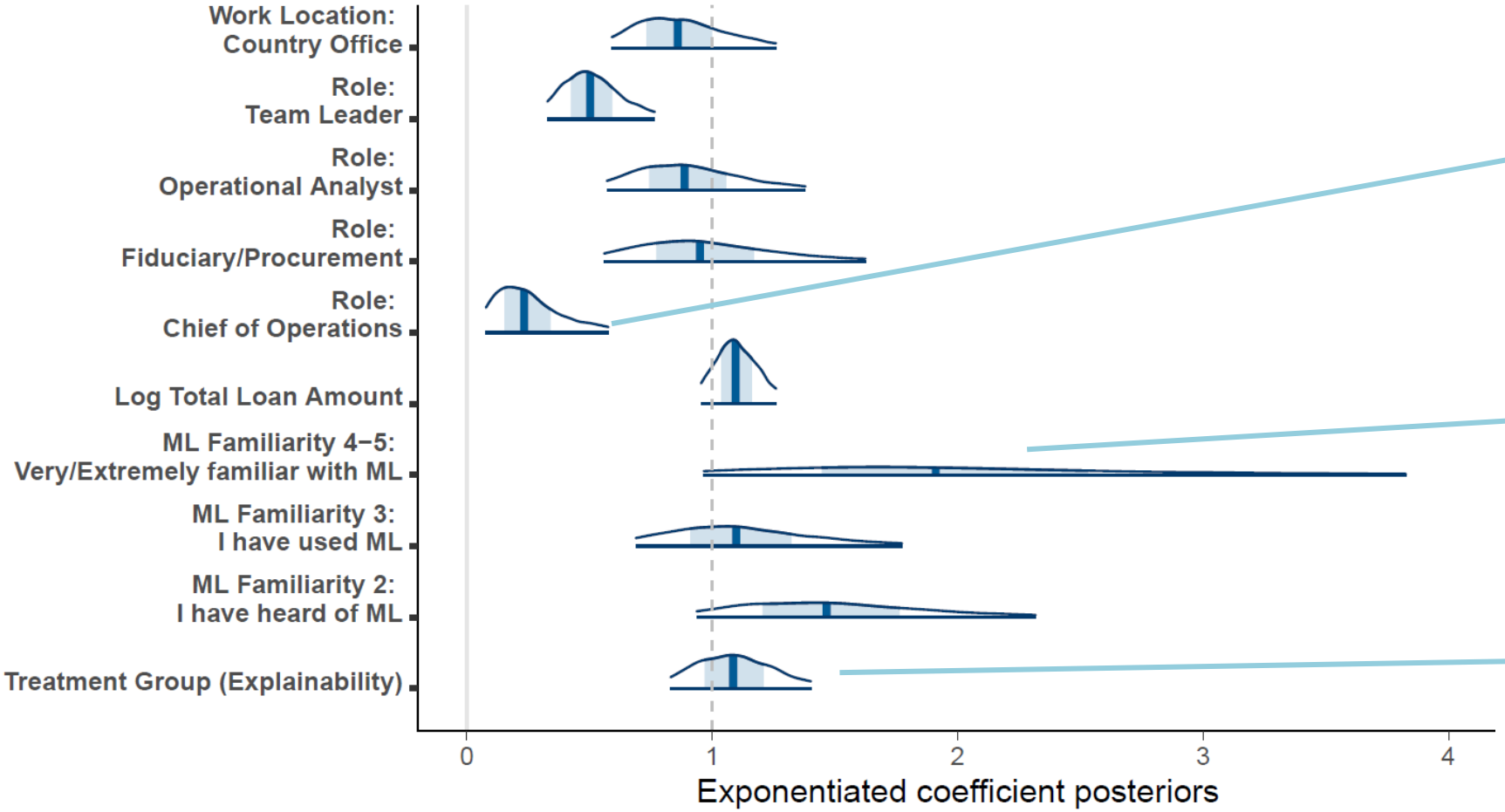
- Most senior member of team updates beliefs by lowest amount

- Higher loan amount -> more likely to update beliefs

- After controlling for other individual characteristics, treatment effect is mixed

Logistic model: After viewing the ML tool did you update your delay estimate?

Outcome: =1 if changed delay estimate for project after viewing ML tool

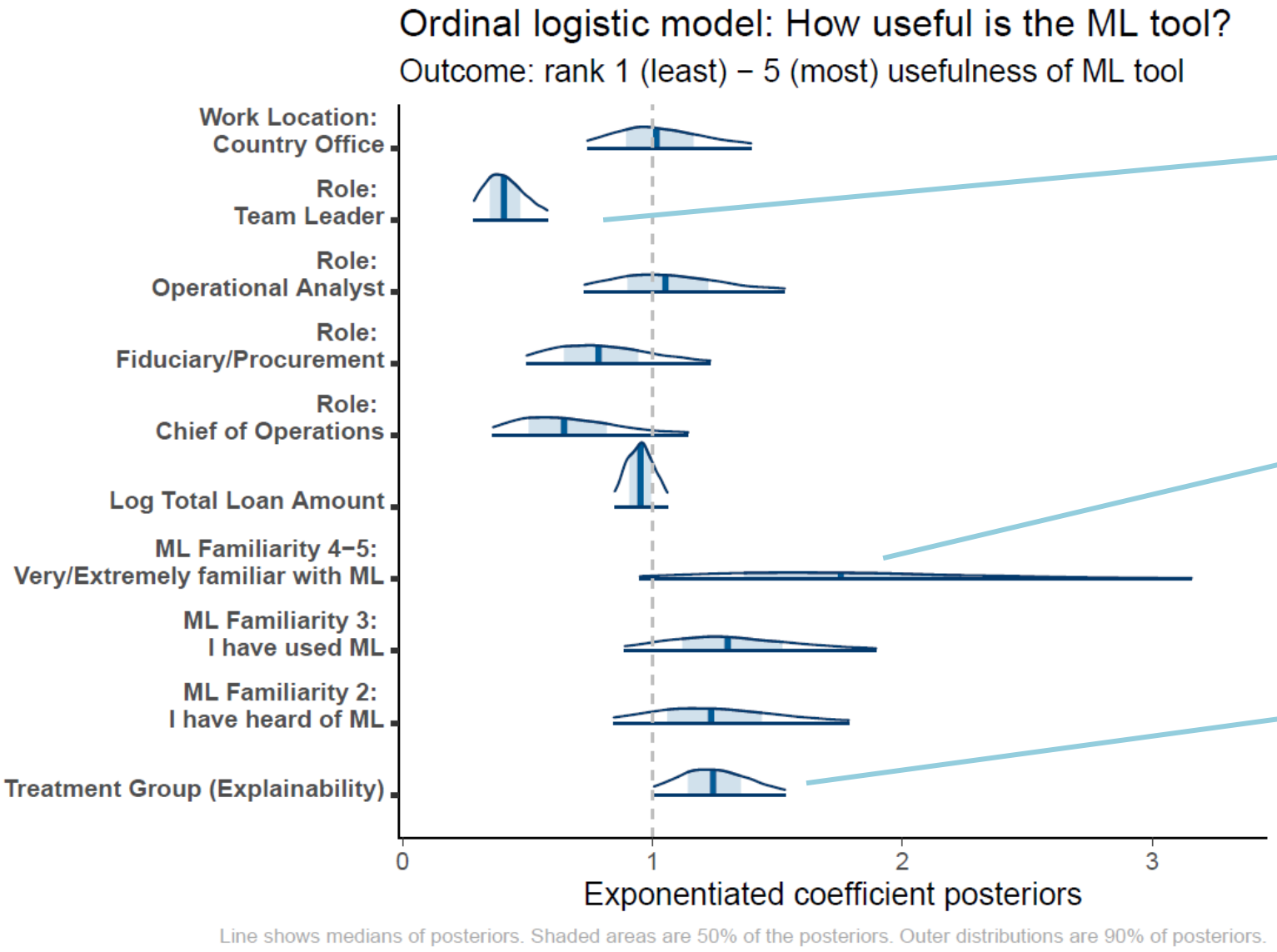


Line shows medians of posteriors. Shaded areas are 50% of the posteriors. Outer distributions are 90% of posteriors.

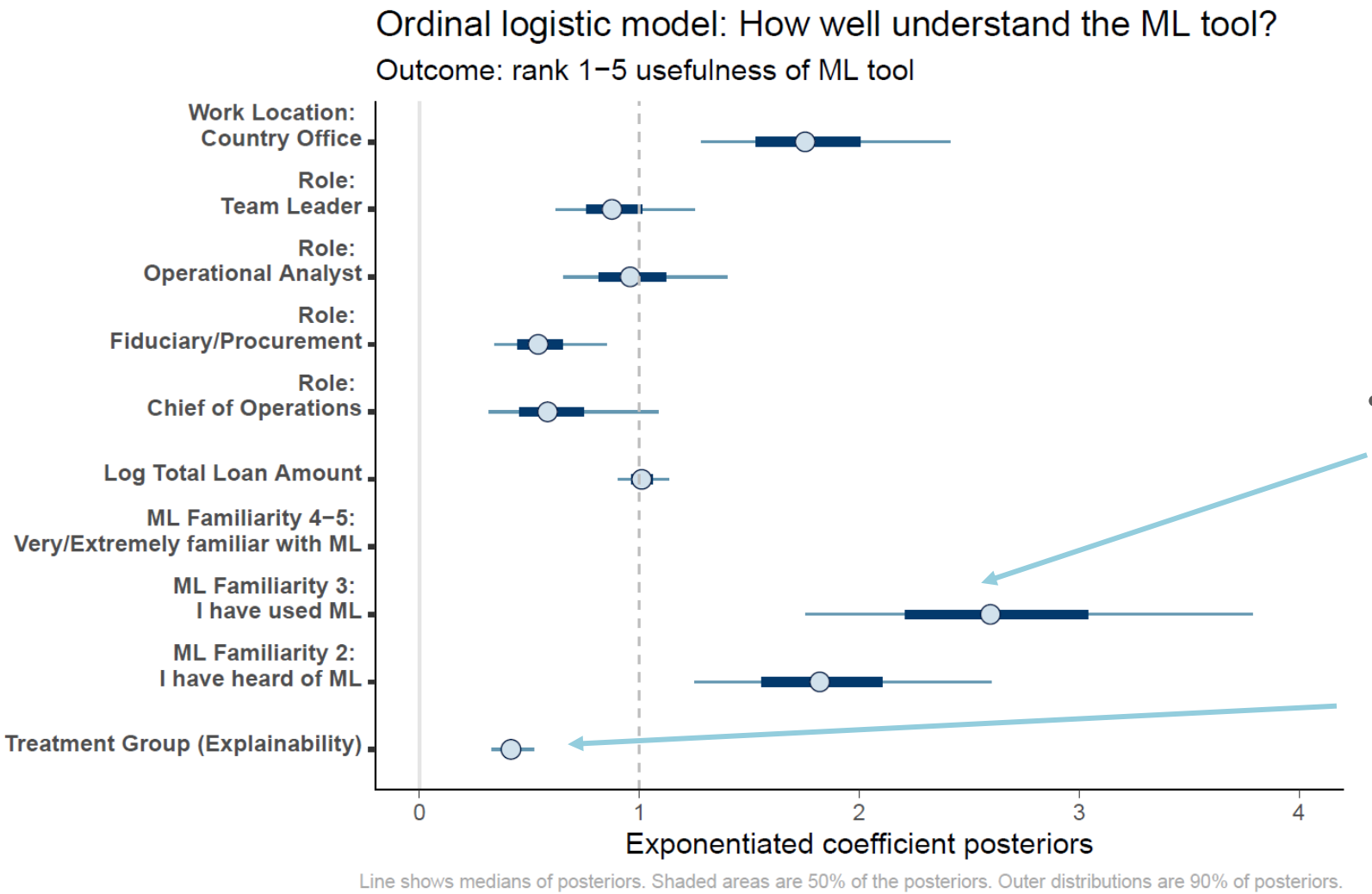
- Median: $\exp(\beta_{Chief}) = 0.23$.
  - Most senior member of team 77% less likely to update delay estimate!

- Most familiar with ML 2x as likely to update delay estimate

- Again, appear to be mixed impact of explainability treatment on belief update



- Direct managers least likely to find the tool useful
- Users with higher machine learning knowledge more likely to find ML tool useful
- Median: $\exp(\beta_{Treatment}) = 1.24$
  ~= 24% increase in likelihood of increasing self reported usefulness if receive explainability treatment



- Self reported ML familiar groups more likely to report understanding tool
- Explainable AI seems to decrease self reported understanding

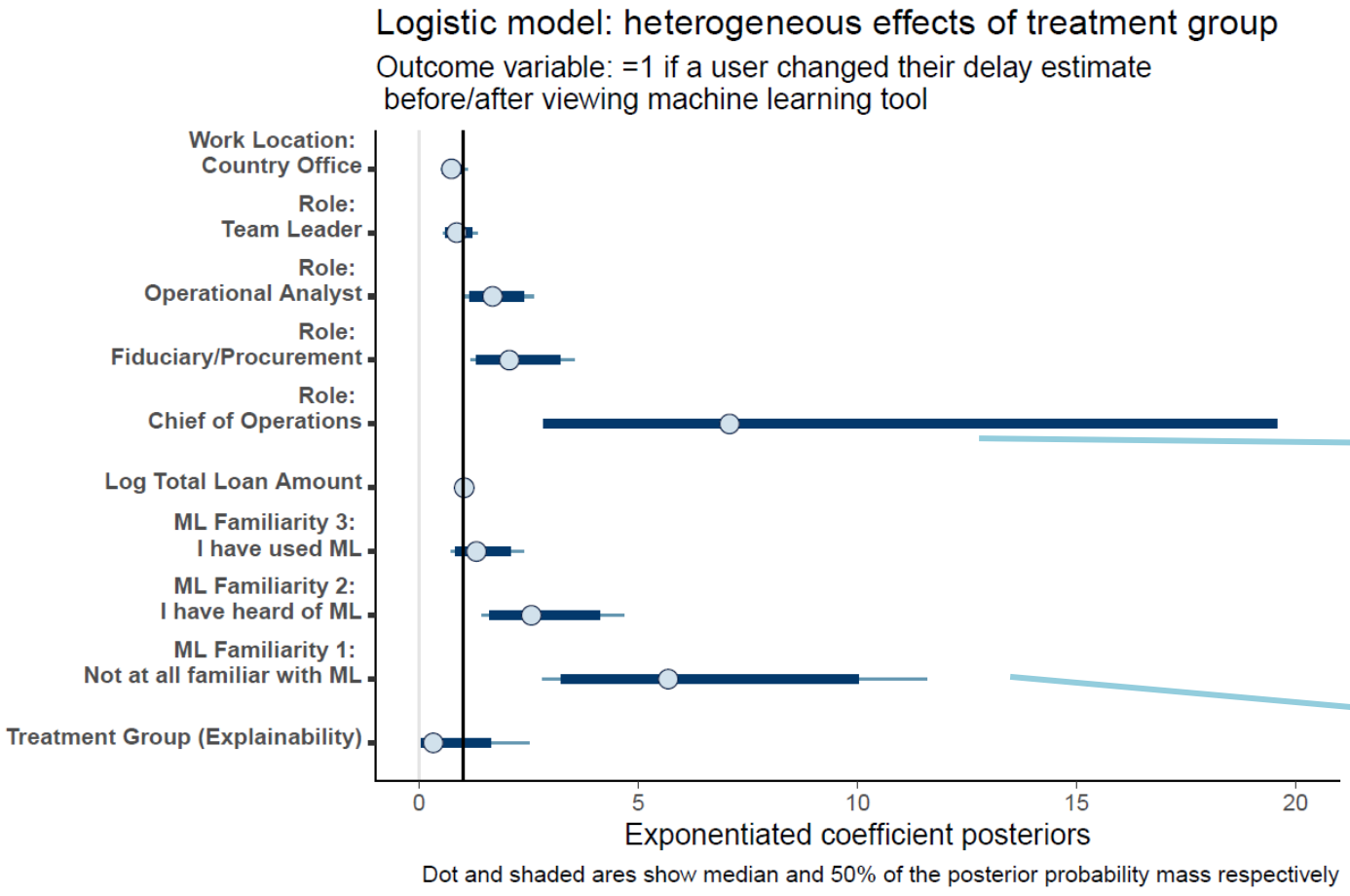
Summary of Results Thus Far

1. Change in their delay estimate before/after viewing tool
2. Absolute value of change in delay estimate update
    - Some evidence explainable AI affects propensity to update or amount to update
    - AI reluctant groups (less ML familiarity, more senior) less likely to update
    - Higher loan amounts -> more likely to update

3. How useful is the ML tool? (Ordered rank 1-5)
    - Evidence Explainable AI increases perceived usefulness
    - More ML familiar groups find more useful

4. How well did you understand the ML tool? (Ordered rank 1-5)
    - Explainable AI decreases
    - More ML familiar groups increases

Which groups are most likely to respond to explainable AI?



Heterogeneous effect of treatment:

$$\sigma(\sum \beta_{Treatment} * x_{i_{moderators}})$$

- Least familiar with ML and most senior members respond most to explainability treatment
- Heterogeneous treatment effects on these groups are very large. 5x-7x more likely to update belief if in explainability treatment

Conclusions

- We find that explainable AI models increases belief updating by 4.5%
- Explainable models increase perceived usefulness but decrease understanding
- Largest loans more likely to update beliefs given ML predictions
- Senior members of team and those least familiar with ML least like to trust AI
- But: explainable models increase belief updating by these reticent groups by 5-7x.
    - Believe this is evidence that some AI reskilling will happen organically as tools improve if they are designed with neophytes in mind

Comments, Questions or Suggestions

People

- Selina Carter, Carnegie Mellon University – shcarter@andrew.cmu.edu
- Jonathan Hersh, Argyros School of Business and Economics, Chapman University – hersh@chapman.edu

Comments, Questions or Suggestions

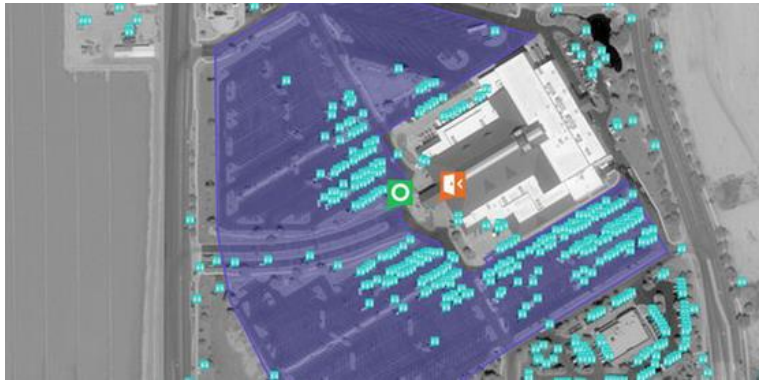
People

- Selina Carter, Carnegie Mellon University – shcarter@andrew.cmu.edu
- Jonathan Hersh, Argyros School of Business and Economics, Chapman University – hersh@chapman.edu

- Assistant Professor, Chapman University Argyros School of Business
- Education: PhD in economics, Boston University, M.Sc. Wharton, BA U Chicago
- Work experience: Data Scientist for the World Bank, economic consultant, and developer
- Research: Economics of information systems, computer vision and applied machine learning,

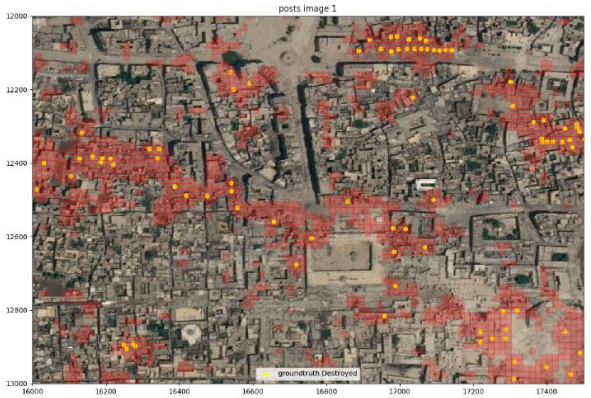
Research Overview

- Satellite Imagery + Computer Vision + Machine Learning



Count cars in parking lots!

Dense Prediction: Scanning Aleppo



Damaged buildings in Syria!

PNAS Proceedings of the National Academy of Sciences of the United States of America

Keyword, Author, or

Home | Articles | Front Matter | News | Podcasts | Authors

RESEARCH ARTICLE

Monitoring war destruction from space using machine learning

Hannes Mueller, Andre Groeger, Jonathan Hersh, Andrea Matranga, and Joan Serrat

+ See all authors and affiliations

PNAS June 8, 2021 118 (23) e2025400118; https://doi.org/10.1073/pnas.2025400118

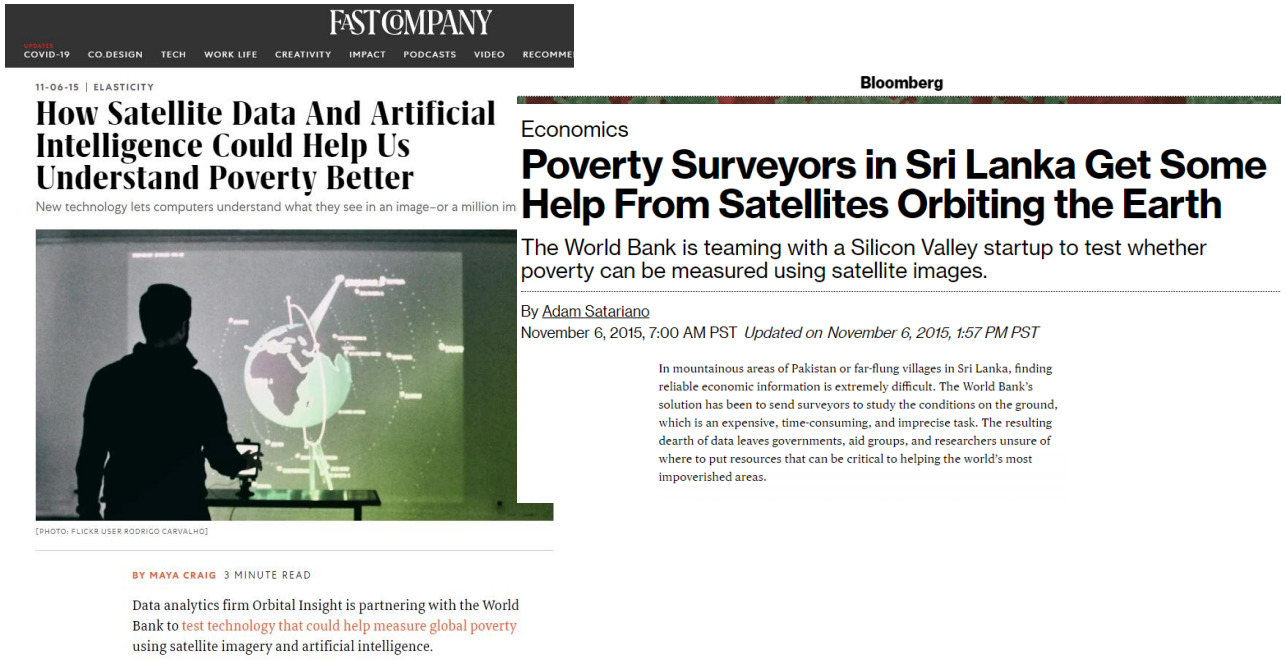
Edited by Douglas S. Massey, Princeton University, Princeton, NJ, and approved April 26, 2021 (received for review December 11, 2020)

Article | Figures & SI | Info & Metrics | PDF

https://www.pnas.org/content/118/23/e2025400118

Research Overview

- Calculating Poverty Using Satellites
- Advised World Bank/IDB on COVID poverty transfers (using methods in this course!) in Belize, Togo, Guinea



FAST COMPANY

COVID-19 CO.DESIGN TECH WORK LIFE CREATIVITY IMPACT PODCASTS VIDEO RECOMME

11-06-15 | ELASTICITY

How Satellite Data And Artificial Intelligence Could Help Us Understand Poverty Better

New technology lets computers understand what they see in an image–or a million im

[PHOTO: FLICKR USER RODRIGO CARVALHO]

BY MAYA CRAIG 3 MINUTE READ

Data analytics firm Orbital Insight is partnering with the World Bank to test technology that could help measure global poverty using satellite imagery and artificial intelligence.

Bloomberg

Economics

Poverty Surveyors in Sri Lanka Get Some Help From Satellites Orbiting the Earth

The World Bank is teaming with a Silicon Valley startup to test whether poverty can be measured using satellite images.

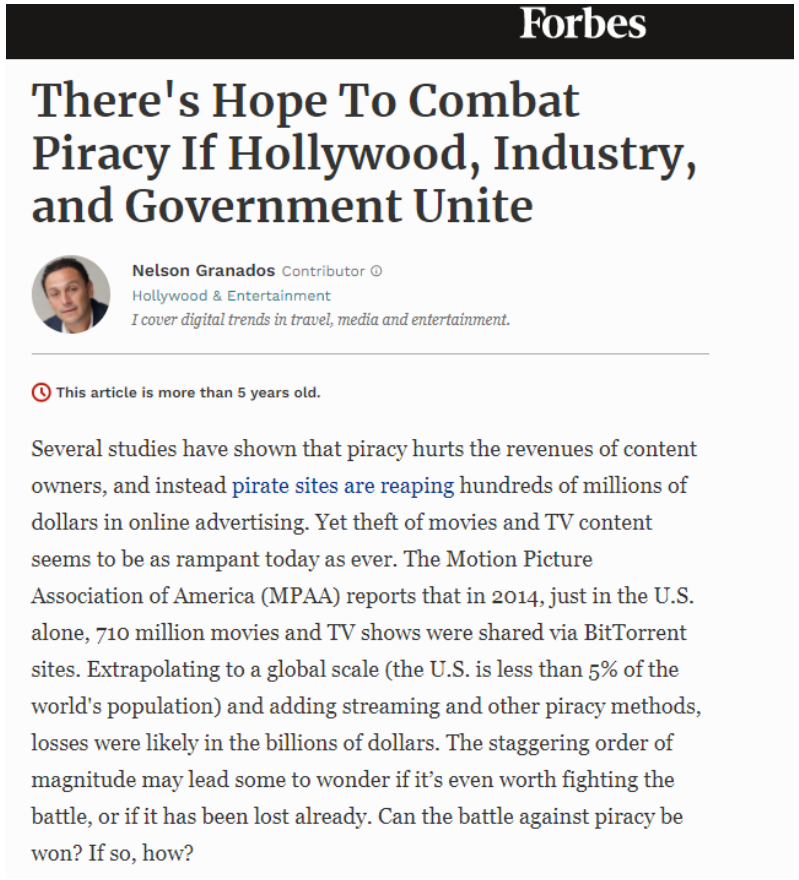
By Adam Satariano

November 6, 2015, 7:00 AM PST *Updated on November 6, 2015, 1:57 PM PST*

In mountainous areas of Pakistan or far-flung villages in Sri Lanka, finding reliable economic information is extremely difficult. The World Bank's solution has been to send surveyors to study the conditions on the ground, which is an expensive, time-consuming, and imprecise task. The resulting dearth of data leaves governments, aid groups, and researchers unsure of where to put resources that can be critical to helping the world's most impoverished areas.

Research Overview 2

- Online Media Piracy



Forbes

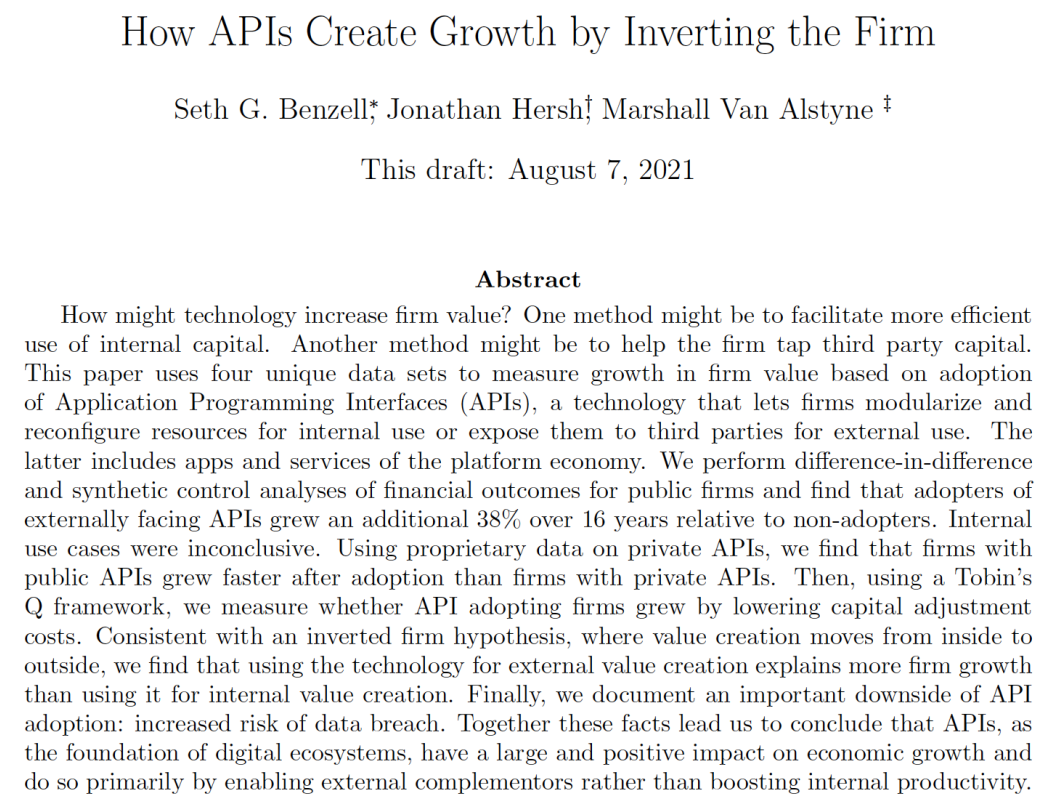
There's Hope To Combat Piracy If Hollywood, Industry, and Government Unite

Nelson Granados Contributor
Hollywood & Entertainment
I cover digital trends in travel, media and entertainment.

This article is more than 5 years old.

Several studies have shown that piracy hurts the revenues of content owners, and instead pirate sites are reaping hundreds of millions of dollars in online advertising. Yet theft of movies and TV content seems to be as rampant today as ever. The Motion Picture Association of America (MPAA) reports that in 2014, just in the U.S. alone, 710 million movies and TV shows were shared via BitTorrent sites. Extrapolating to a global scale (the U.S. is less than 5% of the world's population) and adding streaming and other piracy methods, losses were likely in the billions of dollars. The staggering order of magnitude may lead some to wonder if it's even worth fighting the battle, or if it has been lost already. Can the battle against piracy be won? If so, how?

- Firm IT Strategy



How APIs Create Growth by Inverting the Firm

Seth G. Benzell*, Jonathan Hersh†, Marshall Van Alstyne ‡

This draft: August 7, 2021

Abstract

How might technology increase firm value? One method might be to facilitate more efficient use of internal capital. Another method might be to help the firm tap third party capital. This paper uses four unique data sets to measure growth in firm value based on adoption of Application Programming Interfaces (APIs), a technology that lets firms modularize and reconfigure resources for internal use or expose them to third parties for external use. The latter includes apps and services of the platform economy. We perform difference-in-difference and synthetic control analyses of financial outcomes for public firms and find that adopters of externally facing APIs grew an additional 38% over 16 years relative to non-adopters. Internal use cases were inconclusive. Using proprietary data on private APIs, we find that firms with public APIs grew faster after adoption than firms with private APIs. Then, using a Tobin's Q framework, we measure whether API adopting firms grew by lowering capital adjustment costs. Consistent with an inverted firm hypothesis, where value creation moves from inside to outside, we find that using the technology for external value creation explains more firm growth than using it for internal value creation. Finally, we document an important downside of API adoption: increased risk of data breach. Together these facts lead us to conclude that APIs, as the foundation of digital ecosystems, have a large and positive impact on economic growth and do so primarily by enabling external complementors rather than boosting internal productivity.

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3432591

Most Proud of: Cited on the Wikipedia Page for "Waffle"

Not logged in Talk Contributions Create account Log in

Article Talk | Read View source View history | Search Wikipedia

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Current events
Random article
About Wikipedia
Contact us
Donate

Contribute
Help
Community portal
Recent changes
Upload file

Tools
What links here
Related changes
Special pages
Permanent link
Page information
Cite this page
Wikidata item

Print/export
Download as PDF
Printable version

Waffle

From Wikipedia, the free encyclopedia

This article is about the batter/dough-based food. For other uses, see Waffle (disambiguation).

A **waffle** is a dish made from leavened batter or dough that is cooked between two plates that are patterned to give a characteristic size, shape, and surface impression. There are many variations based on the type of waffle iron and recipe used. Waffles are eaten throughout the world, particularly in Belgium, which has over a dozen regional varieties.[1] Waffles may be made fresh or simply heated after having been commercially cooked and frozen.

Waffle

Contents [hide]
1 Etymology
2 History
2.1 Medieval origins
2.2 14th–16th centuries
2.3 17th–18th centuries
2.4 19th–21st centuries
3 Varieties
4 Toppings
5 Consistency
6 Shelf stability and staling
7 See also

Place of origin	France, Belgium
Main ingredients	Batter or dough
Variations	Liège waffle, Brussels Waffle, Flemish Waffle, Bergische waffle, Stroopwafel and others

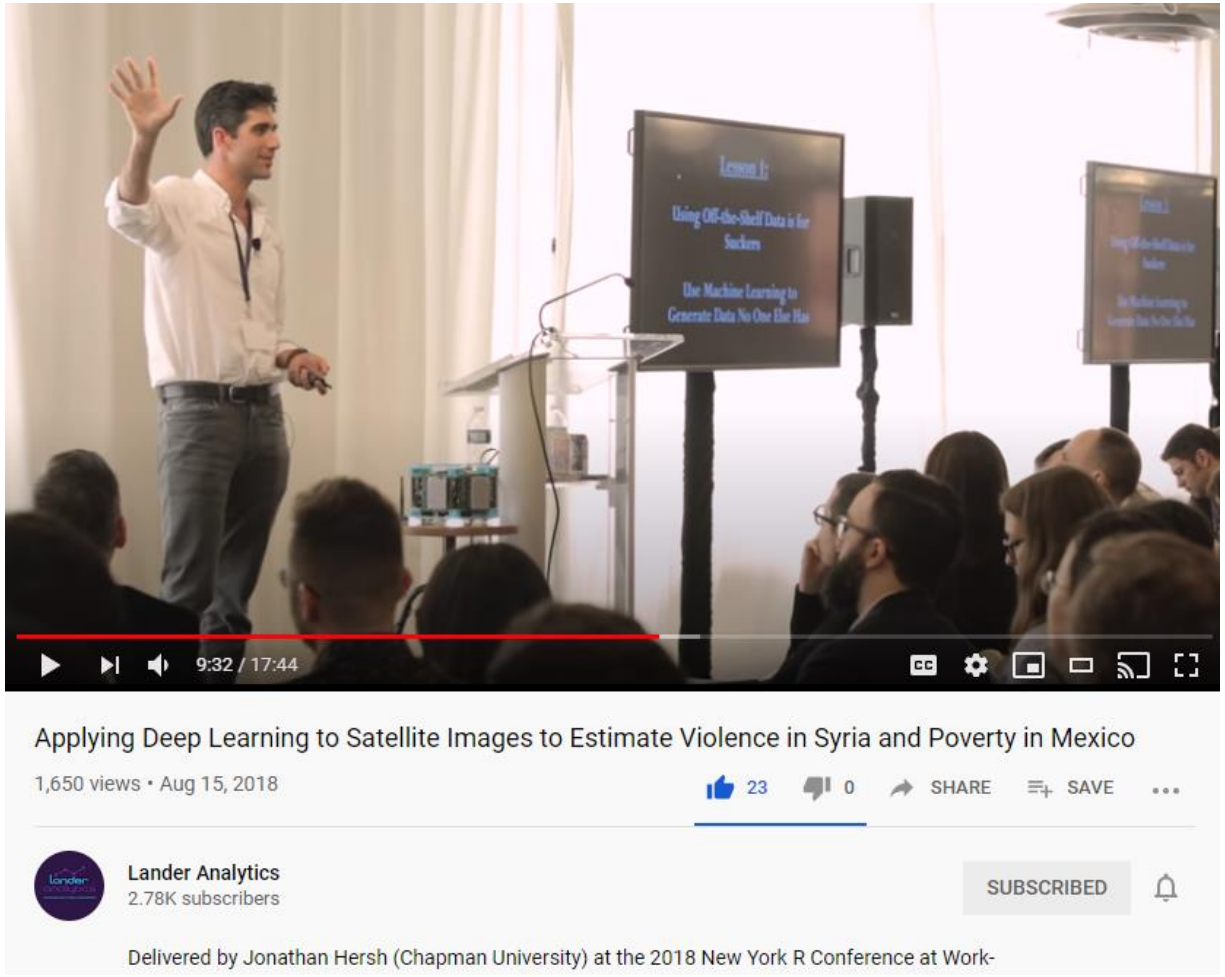
Cookbook: Waffle
Media: Waffle

References

1. ^ "Les Gaufres Belges" Archived 2012-08-20 at the Wayback Machine. Gaufresbelges.com. Retrieved on 2013-04-07.
2. ^ Robert Smith (1725). *Court Cookery*. p. 176.
3. ^ "Waffle" Archived 2013-04-07 at the Wayback Machine, The Merriam-Webster Unabridged Dictionary

52. ^ a b "Sweet Diversity: Overseas Trade and Gains from Variety after 1492" Archived 2013-07-26 at the Wayback Machine, Jonathan Hersh, Hans-Joachim Voth, Real Sugar Prices and Sugar Consumption Per Capita in England, 1600–1850, p.42

I Have Given Talks for the R Community



https://youtu.be/pLqL7qhi6pw

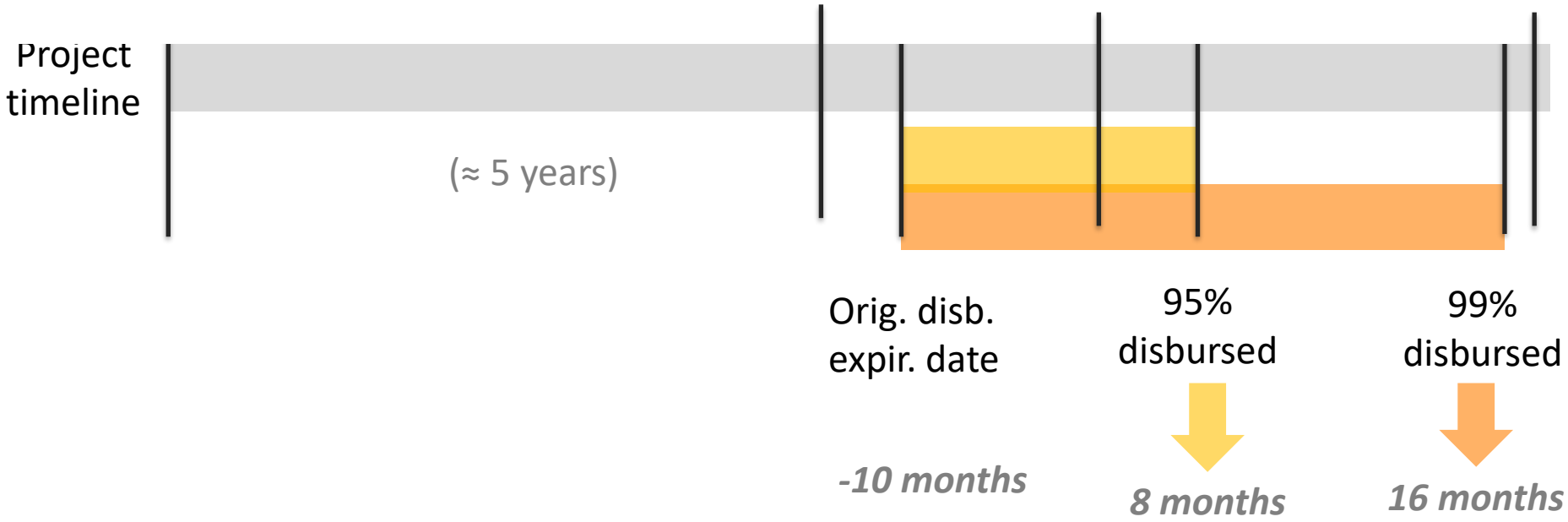
What's the **data**?

Universe considered

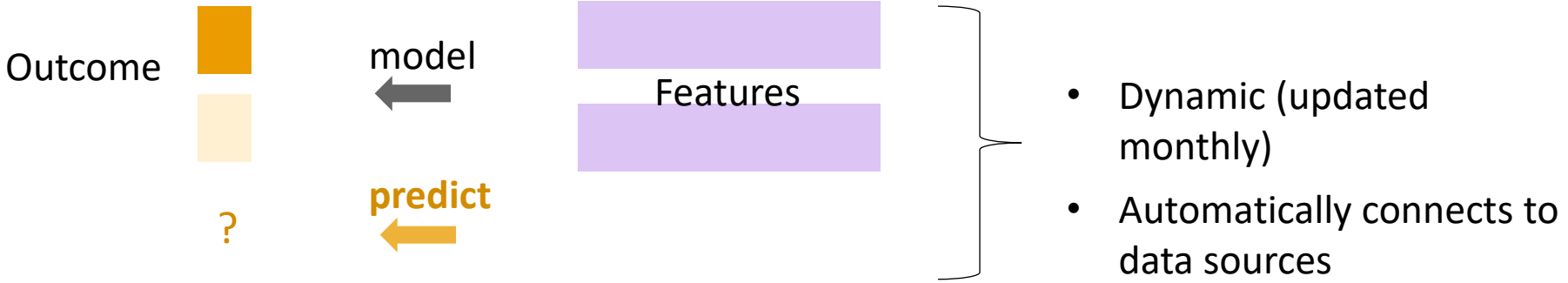
- SG investment loans
- Approved since 2000
- Have an "original disbursement expiration date"

Outcome variable

- Time from **95%** **or** **99%** **of funds disbursed** to **original disbursement expiration date**



What's the **data**?



- Dynamic (updated monthly)
- Automatically connects to data sources

Project age (months since approval)	Project	Outcome (months delayed)	Feature 1	Feature 2
0	AR-L1001	45.9	abc	.2
1	AR-L1001	45.9	abc	.4
2	AR-L1001	45.9	abc	.4
3	AR-L1001	45.9	abc	.5
4	AR-L1001	45.9	abc	.6

What's the data?

100+ features

(Fixed over time)

(Dynamic)

Fixed variables

- Country
- Department
- Approval year
- Approval quarter
- Approved amount ($)
- ESG classification
- Modality
- Approval procedure
- Country requires ratification

Preparation data

- Time spans (months)
  - On Convergence → Start
  - Start → ERM
  - … etc.
  - Approval → Eligiblity
  - Eligibility → First disbursement
- Number of missing date fields in Convergence
- Total cost of preparation (PC + NPC)
- Time and Labor in preparation

Relationship data

- Has OPC TC
- Part of credit line
- Part of sequence
- # in sequence
- # of loan contracts

Team Leader data

- # changes in TL
- # projects TL managing
- # projects has managed in past
- Years been a TL
- *(no personal data)*

Executive agency experience

- Type of agency (ministry, municipality, etc.)
- Driving distance (km and minutes) to COF
- # of projects agency simultaneously managing
- # of projects has managed in past
- Years been a client of IDB

Time snapshot

- Year of snapshot
- Time → deadline
- Time → current disbursement expiration date

Findings & recommendations

- language
- # of characters used
- # of fields entered
- Keyword search: "delay" and "disbursement"

Disbursements

- % disbursed

Results data

- PMR stage
- CPI & SPI
- Synthetic indicator (PMR)

Additional variables

- External country-level variables (GDP growth, etc.)
- More time & labor info
- Expenses ($) data
- Team composition
- Ideas are welcome!