IRS

Research, Applied Analytics, Statistics (RAAS)

Integrating Trustworthy Analytics

Jameson Carter, Chloe Zheng, Brian Xu

August 19, 2022

Our Team

Research, Applied Analytics, & Statistics (RAAS)

- Stephanie Needham, *Project Lead*
- Holly Donnelly, *Interim Project Lead*
- Alissa Graff, *Tools Workstream Lead*
- Frank Cousin, *Training Workstream Lead*
- Isaac Schwab
- Yan Sun, *Literature Review Workstream*
- Nathan Gire, *Literature Review Workstream*
- Traci Suiter
- Annelise Britten
- Tom Hertz
- Civic Digital Fellows (Summer 2022)
    - Jameson Carter
    - Chloe Zheng
    - Brian Xu

Equity, Diversity, & Inclusion (EDI)

- Jon Ocana
- Michael Sebastiani

What is Trustworthy Analytics?

Trustworthy Analytics is a manner of designing, developing, acquiring, and using AI tools and analytic tools in Public Algorithmic Systems in a way which **fosters public trust and confidence** while protecting privacy, civil rights, and civil liberties.

[HHS Trustworthy AI Playbook, p. 5, Pittsburgh Task Force on Public Algorithms, p. 8]

Project

Objective:

The objective of this research is to mitigate risk by recommending best practices form incorporating trustworthy analytics into RAAS analytic processes.

Research will include:

- Gaining an understanding of current private and public sector best practices.
- Identifying necessary data, testing tools, and methods.
- Increasing awareness throughout RAAS.

FY22 Activities:

- **Conduct a literature review** to identify trustworthy analytics best practices and develop a trustworthy analytics toolkit to share best practices.
- **Conduct a pilot** to assess data, tools, and methods, and develop recommendations that should be considered and applied when developing and implementing analytic solutions.
- **Develop training solutions** on trustworthy analytics and deliver them to the RAAS Community.

Chloe: Overview of Principles

What does this cover?

We provide an overview of six principles that are imperative to foster and strengthen trust in analytic systems.

Fairness and Bias	Transparency, Interpretability, Explainability	Reliability
Privacy	Security	Accountability

Why is this important?

As automation and artificial intelligence systems gain popularity, analytic teams must understand these principles – these ideas and concepts are recognized across international agencies, U.S. agencies, the private sector, and academic institutions. The overview covers risks and case studies that demonstrate consequences when teams fail to satisfy these principles.

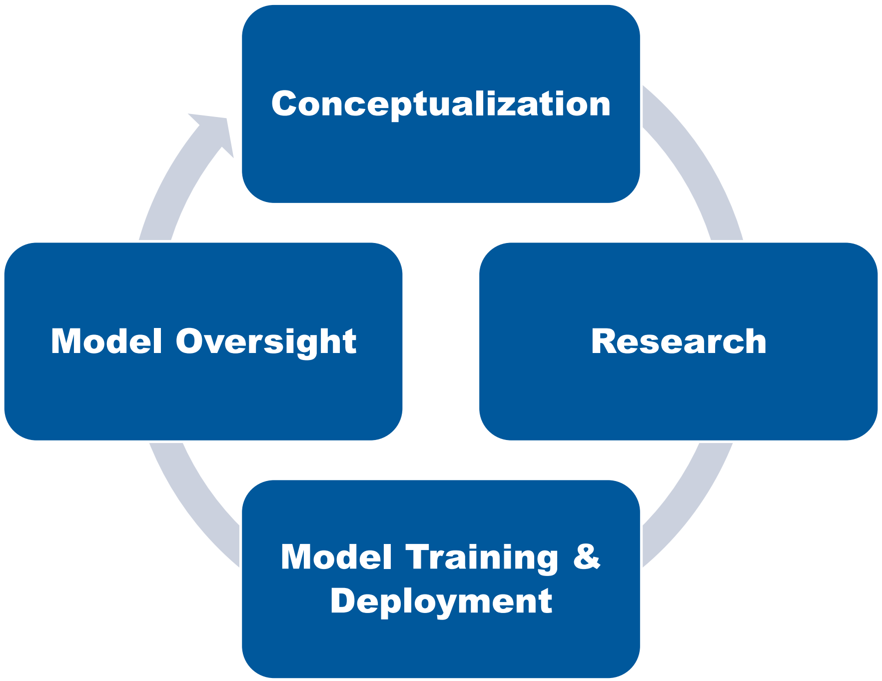
It is in the IRS's best interest that their analysts and management teams are up to date on important principles and practices to promote trust and mitigate risk.

Jameson: Framework for teams

Teams looking to pursue Trustworthy Analytics now have a framework they can follow which:

- Provides **bias mitigation checklists** for each stage of the analytics lifecycle.
- Gives a broad view of how to **design fair systems.**
- Helps teams **affirmatively plan fairness** into systems.

The Analytics Lifecycle



Brian: Toolkit for Analysts

The **analyst toolkit** provides actionable knowledge and steps for analysts to take when assessing fairness and mitigating bias within their models and algorithms.

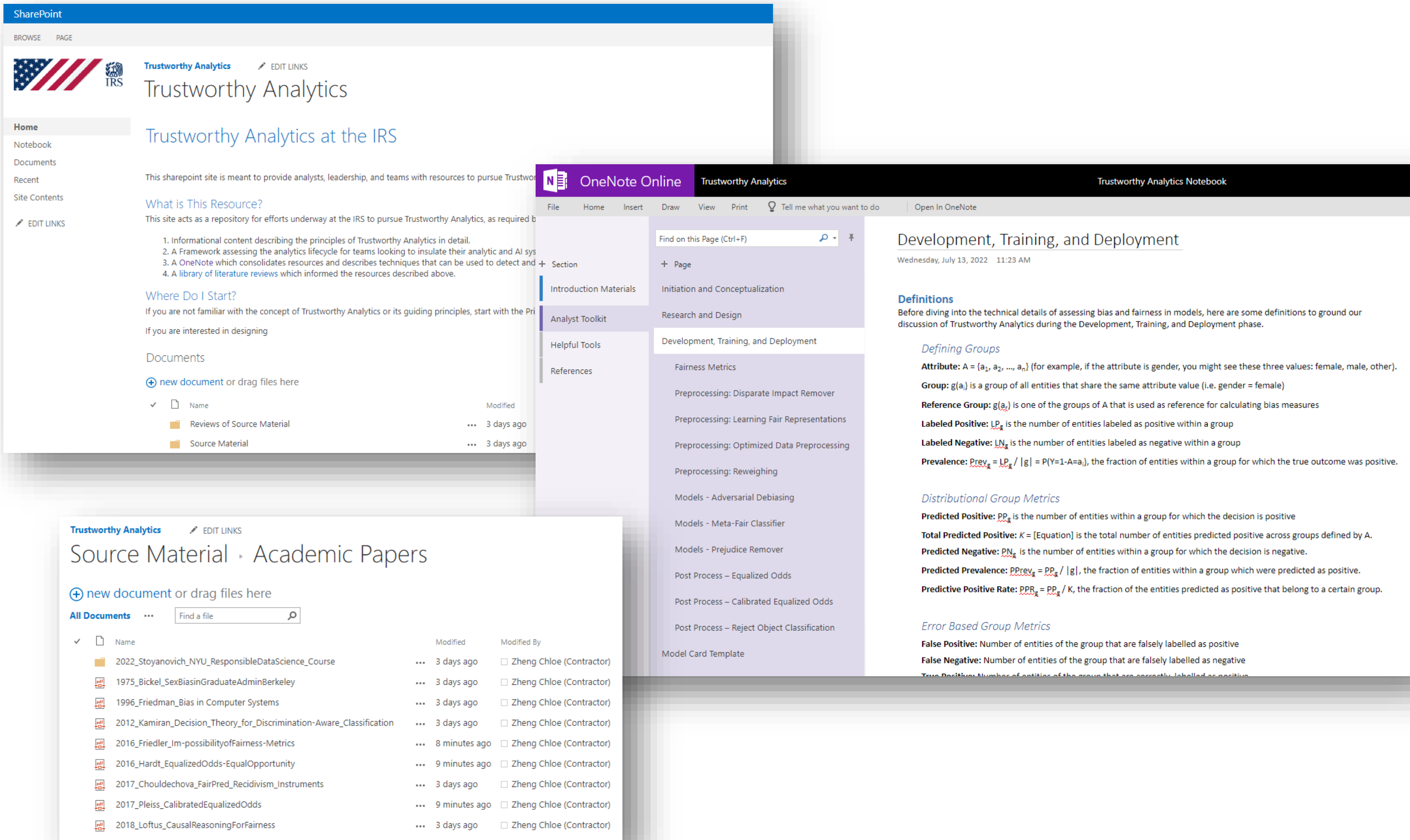
Key Considerations
- **Risk classifications**
- **Project scoping**
- **Consulting stakeholders**

Fairness Metrics
- **Defining groups**
- **Defining fairness**
- **Quantifiable**

Bias Mitigation
- **Pre-processing**
- **In-processing**
- **Post-processing**

The goal of the toolkit is to provide analysts with a solid understanding of **points to consider** when assessing and developing models, as well as provide **tools for increasing fairness** in models.

Trustworthy Analytics SharePoint



Next Steps

- **Literature Review**
  - Refine SharePoint Site and existing guidelines and toolkits
  - Continue to collect and organize resources – expand on topics that were not covered
  - Utilize resources to assist Training Workstream
- **Toolkit Pilot**
  - Assess existing metrics and algorithms to understand best use cases
  - Develop tools training
  - Full implementation and integration of relevant and necessary tools for RAAS Analysts