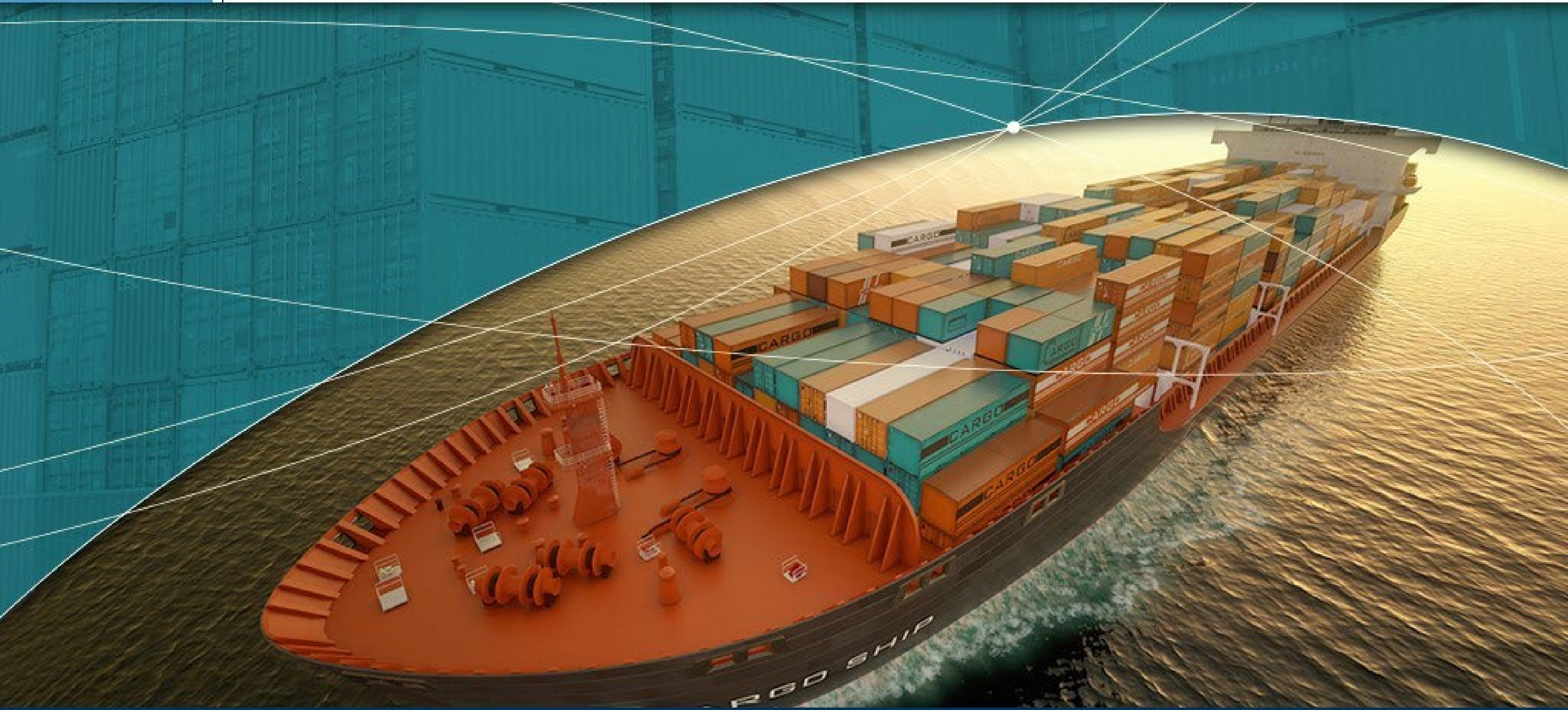
Securing Confidentiality in Machine Learning Models

Ellen Galantucci



Confidentiality and AI

- Increasing concerns about how AI methods expose government data to risk
- Many federal statistical agencies are changing disclosure limitation methods as a result
- But what about the risks posed by government employees using AI to produce/publish federal data?

Creating AI Models for Internal Use

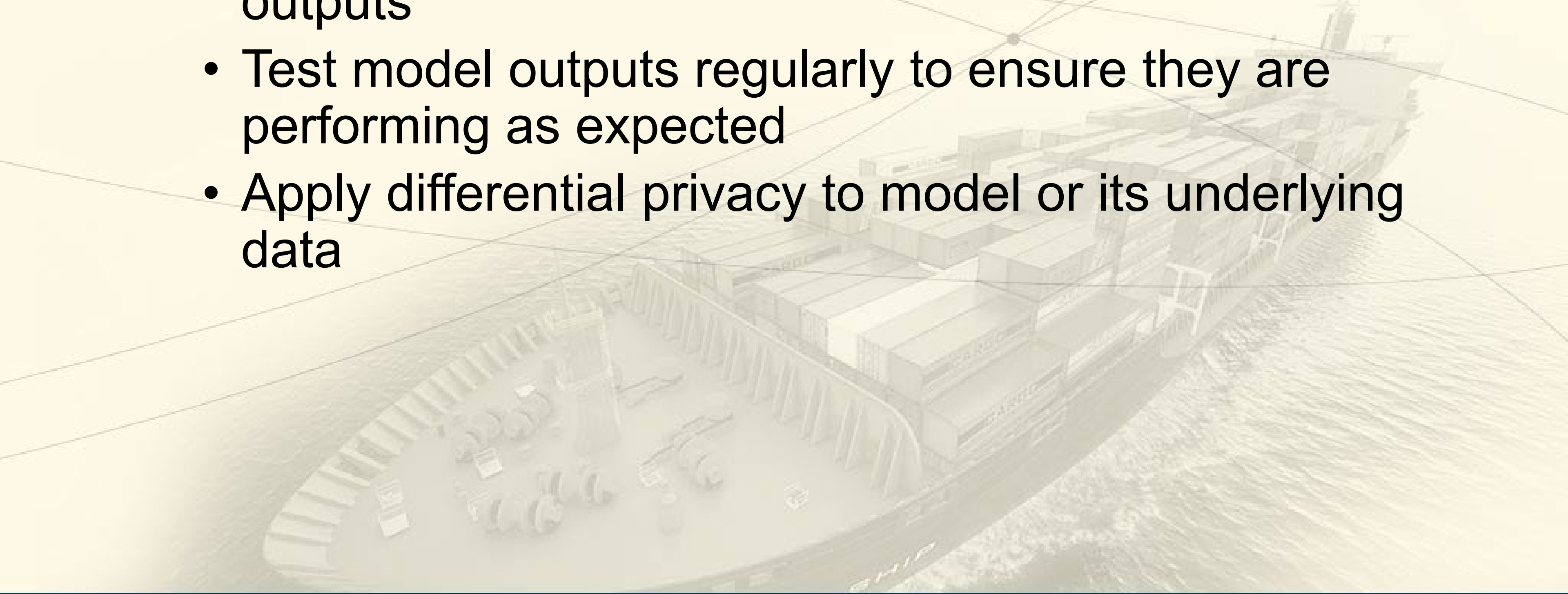
- Confidential data are often used to create internal AI models
- Examples include:
  - A natural language processing (NLP) autocoder
  - Image classifiers

Risks in Internal Model Creation

- AI models can produce biased outputs that inadvertently expose confidential information
- Example risks include:
    - Revealing sensitive data through unexpected patterns
    - Improper handling of small or specific subsets of the training data

Mitigating Risks in Internal Models

- Preventing disclosure risk:
  - Ensure training data reflect broader population outputs
  - Test model outputs regularly to ensure they are performing as expected
  - Apply differential privacy to model or its underlying data



Releasing AI Models to the Public

- Risks of public model release:
  - Models can inadvertently store and expose underlying data
  - Overfitting can lead to the model memorizing sensitive information

Mitigating Risks in Public Model Releases

- Preventing disclosure risk:
  - Limit training data inputs to higher-frequency, less sensitive information
  - Apply differential privacy techniques to protect underlying data
  - Use a controlled, internal process for public interaction with the model (e.g., outsiders submit data to the model for process, without releasing the model itself)

Case Study: Predicting Freight Rates Using AI

- The Federal Maritime Commission (FMC) collects service contracts between shippers and common carriers. These contracts outline the rates paid for transportation services, including detailed information like cargo type, shipping routes, and volume
- Contracts come in many forms and vary in content depending on the agreements between the parties
- The diversity and complexity of the contracts make it difficult to analyze freight rate trends manually

Challenge

- FMC needed to gain insights into freight rate fluctuations across carriers and routes, but:
  - Contracts were often in unstructured text formats
  - There was significant variation in terms of conditions, making manual analysis prone to errors
  - The dataset contained controlled, unclassified information that could not be shared outside of the agency in its raw form

AI Solution

- Natural Language Processing (NLP) was used to create an AI model capable of auto-coding contract data into standardized categories (e.g., service type, destination)
- The model was trained using the most standardized contracts and then expanded to include additional carriers and shippers until it could predict larger portions of the database

Confidentiality Risks

- Model overfitting: The was a risk that the model would memorize specific rates and carriers, potentially exposing confidential contract details
- Biased predictions: The small number of carriers in the dataset could lead to biased freight rate predictions, disproportionately revealing the activities of a few companies

Risk Mitigation Strategies

- Differential privacy: We are exploring implementing a differential privacy strategy to the model outputs, adding noise to the predictions to ensure no individual contract could be reverse-engineered from the results
- Limiting sensitive data: We have also explored excluding contract terms and routes that are niche to the market and could be easily identified

Outcomes

- We are improving our ability to predict market-wide freight rate trends without exposing sensitive contract details
- Privacy preservation: FMC is better able to maintain the confidentiality of the carriers’ data while still providing useful insights into the market trends

Conclusions

- AI models, whether internally created or released publicly, can pose significant disclosure risks
- Careful attention to model design, training, and release procedures can reduce these risks while ensuring the usefulness of AI models in government settings

Contact Information

Ellen Galantucci

Senior Data Scientist

Federal Maritime Commission

egalantucci@fmc.gov