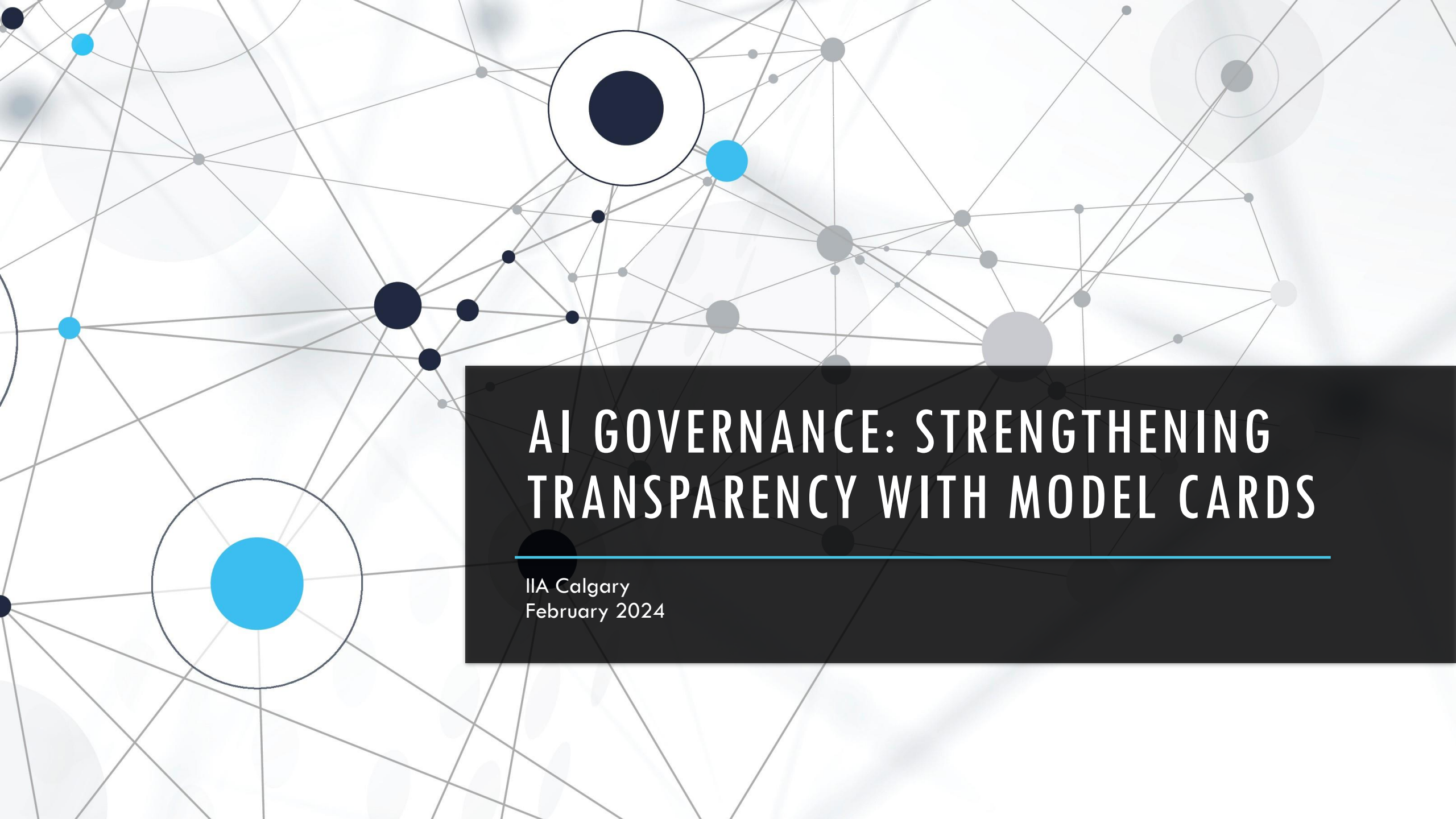


AI GOVERNANCE: STRENGTHENING TRANSPARENCY WITH MODEL CARDS

IIA Calgary
February 2024

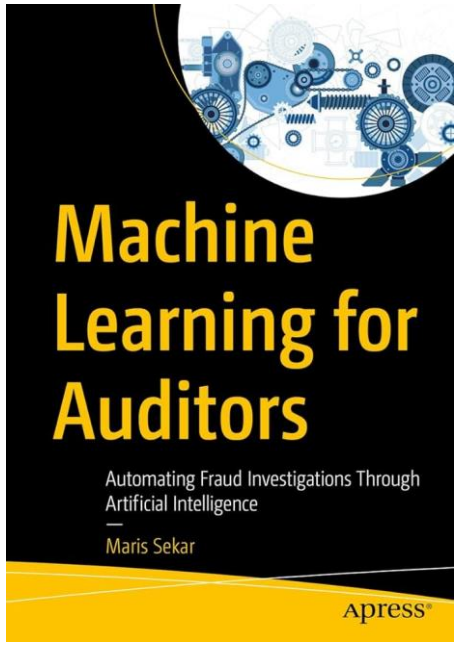
INTRODUCTION

pwc

TC Energy

équité
Defending Against Insurance Crime

AltaML



UNIVERSITY OF CALGARY

MIT PROFESSIONAL EDUCATION

HARVARD UNIVERSITY

IEEE SMC
Systems, Man, and Cybernetics Society

MANNING

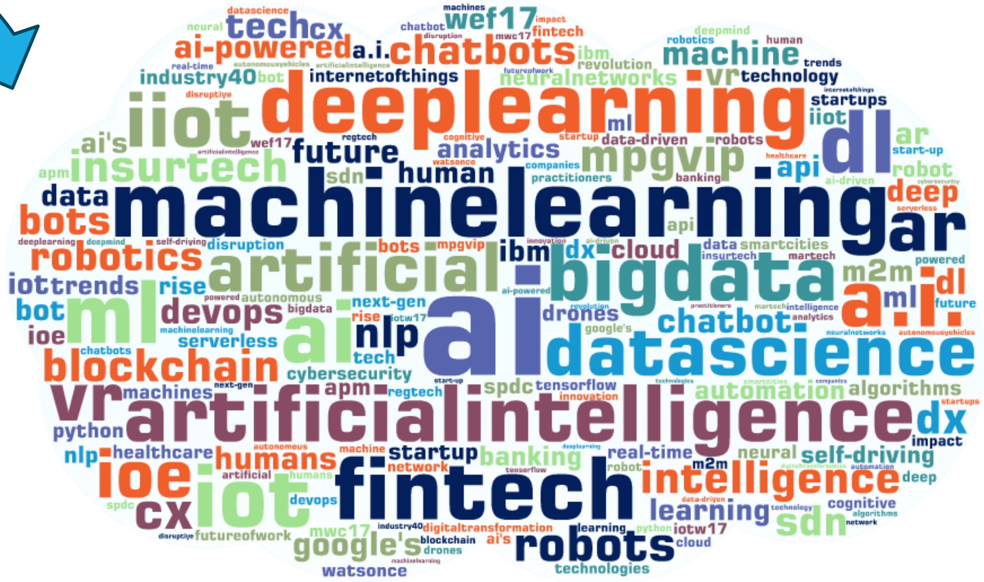
POLLING QUESTION #1

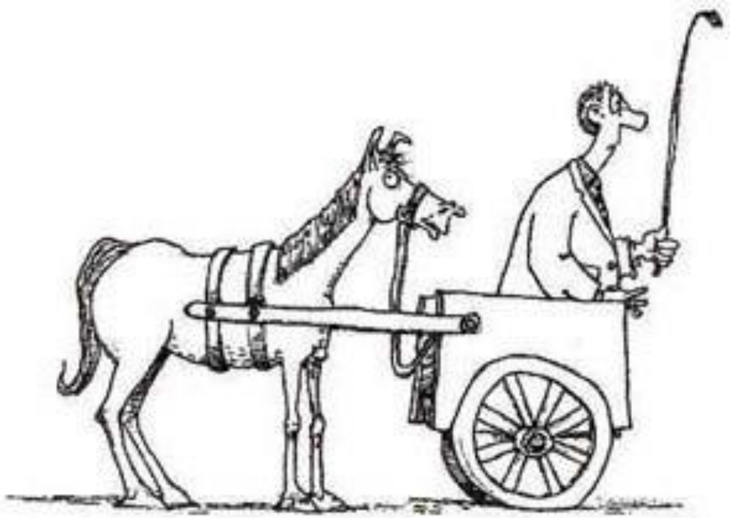
What types of AI projects are prevalent at your company?

- ❑ RPA / Chatbots
- ❑ Text Recognition / Computer Vision
- ❑ Recommenders / Prediction Models
- ❑ No AI Projects

THE PROBLEM


WHERE SHOULD WE FOCUS THIS YEAR?
"BLOCKCHAIN"
IT WILL CHANGE EVERYTHING.
EVERYBODY IS TALKING ABOUT IT.
THE POTENTIAL APPLICATIONS ARE ENDLESS.
WE DON'T WANT TO BE LEFT BEHIND.
WHAT EXACTLY IS BLOCKCHAIN?
ALSO, "ARTIFICIAL INTELLIGENCE"
TOM FISH BURNE
© marketoonist.com


deeplearning
machinelearning
artificialintelligence
bigdata
ai
datascience
fintech
iot
blockchain
artificial
robotics
chatbots
insurtech
iiot
ml
vr
dl
a.i.
intelligence
robots
ioe
cx
devops
nlp
google's
humans
ai-powered
tech
machine
mpgvip
future
human
analytics
chatbot
learning
sdn
ar
bots
data
deep
wef17
industry40
internetofthings
neuralnetworks
technology
startup
banking
self-driving
automation
algorithms
cybersecurity
python
healthcare
serverless
drones
cloud
ibm
watsonce
technologies
futureofwork
digitaltransformation



AUDIT CHALLENGES

- ❑ Fear Of Missing Out (FOMO)
- ❑ Technology-centered vs. Business problem driven
- ❑ Iterative in Nature - Experimental
- ❑ Jargon Madness – “in ENGLISH, please!”

POLLING QUESTION #2

What problem is your company facing with AI projects?

- ❑ Quantifying business value
- ❑ Ethical concerns
- ❑ Data quality
- ❑ Other / Not listed

DO YOU NEED TO KNOW HOW?!

Top Selling Costco Products*:

1. Toilet Paper
2. Rotisserie chicken
3. Bacon

How is Rotisserie Chicken made?

- ❑ Ingredients
- ❑ How is it prepared?
- ❑ Cooking temperature and time?

Nutrition Facts

8 servings per container

Serving size 2/3 cup (55g)

Amount per 2/3 cup	
Calories	**230**
% DV*	
12%	**Total Fat** 8g
5%	Saturated Fat 1g
	Trans Fat 0g
0%	**Cholesterol** 0mg
7%	**Sodium** 160mg
12%	**Total Carbs** 37g
14%	Dietary Fiber 4g
	Sugars 1g
	Added Sugars 0g
	Protein 3g
10%	**Vitamin D** 2mcg
20%	**Calcium** 260mg
45%	**Iron** 8mg
5%	**Potassium** 235mg

* Footnote on Daily Values (DV) and calories reference to be inserted here.

*https://www.theglobeandmail.com/investing/markets/stocks/COST-Q/pressreleases/10574214/here-are-5-of-costcos-most-popular-products

MODEL CARD TEMPLATE

Model Card - Title

Model Details
- Developers
- Model Date, Version & Type
- Training algorithms
- Resources, Citation, License

Evaluation Data
- Details on data used for quantitative analysis
- Datasets, Motivation, Preprocessing

Training Data
- Same detail as evaluation data if possible (privacy constraints)
- Details of distribution over factors

Intended Use
- Primary intended uses & users
- Out of scope use cases

Quantitative Analysis
Unitary & intersectional results

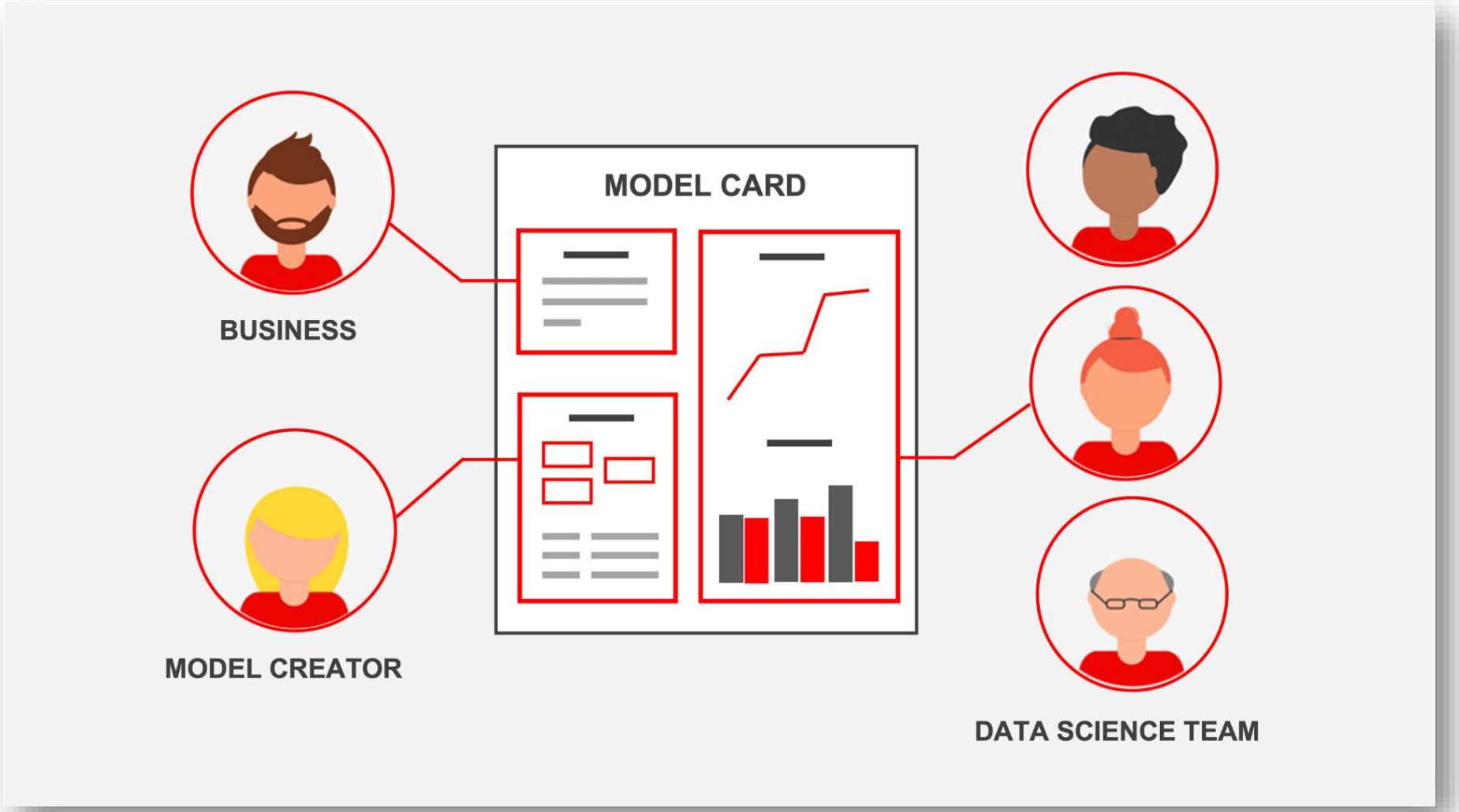
Factors
- Groups, Environments, Instrumentation
- Relevant factors & evaluation factors

Metrics
- Model performance measures
- Decision thresholds
- Variation approaches

Ethical Considerations
- Bias, fairness, ethical considerations
- Mitigation efforts

Caveats, Recommendations
- Concerns not already covered
- Usage information
- Limitations, risks, trade-offs

COLLABORATED DELIVERABLE



https://www.evidentlyai.com/blog/how-wayflyer-creates-ml-model-cards

MODEL CARDS

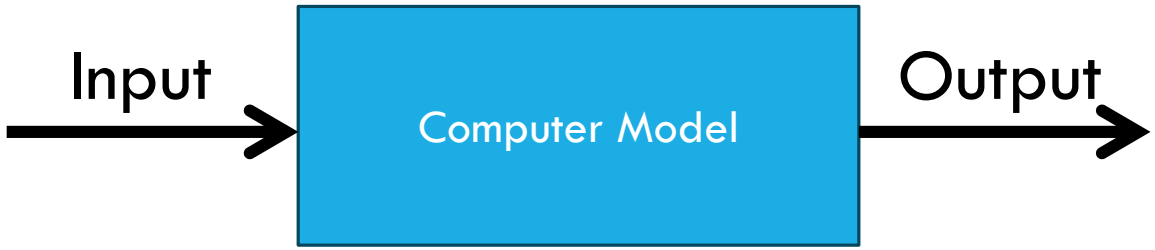
SIMPLE LANGUAGE

UNDERSTAND
BEHAVIOR AND BIASES

CONCISE

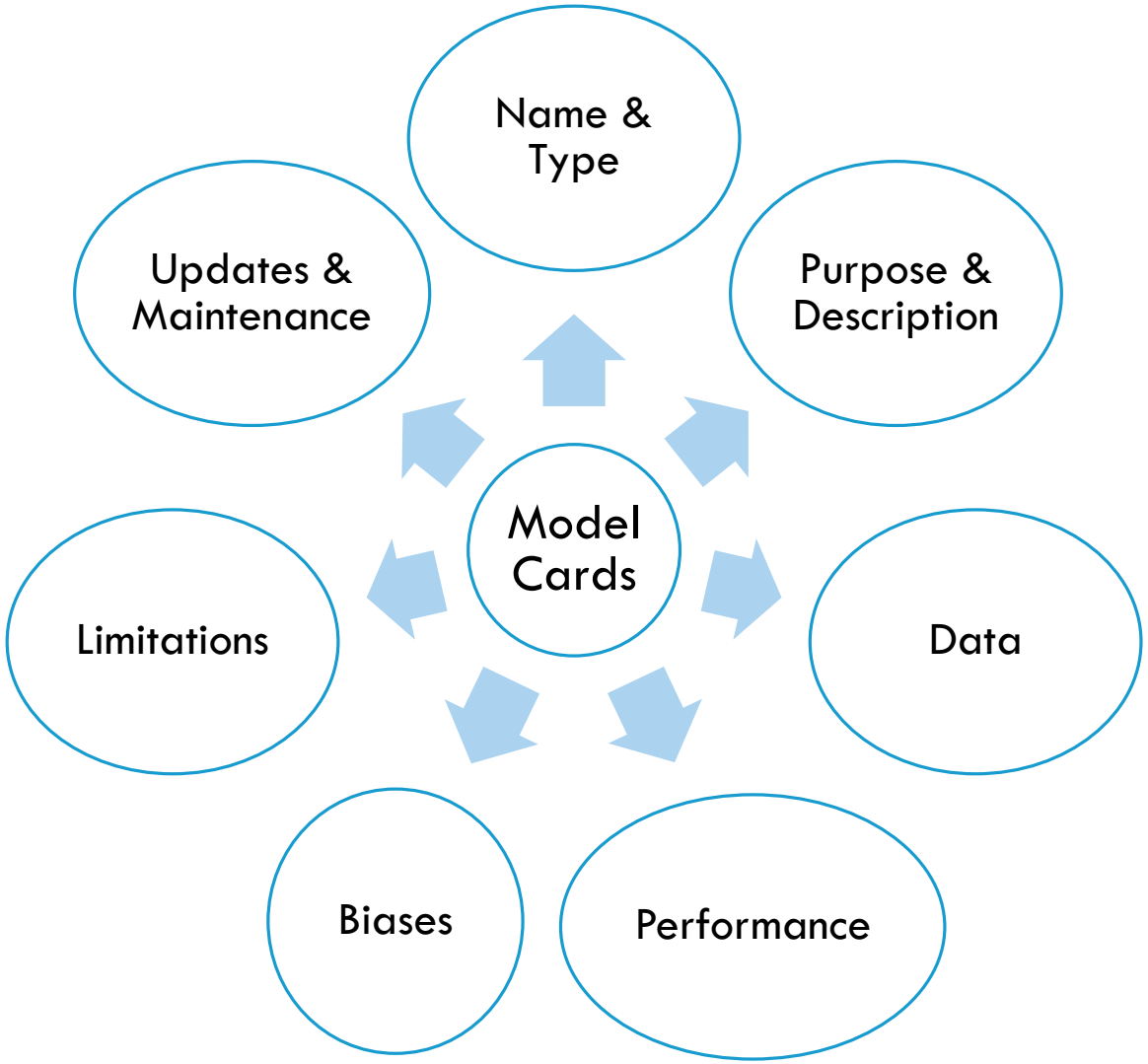
MODEL BEHAVIOR AND BIASES

- How a computer model acts in response to inputs



- Biases can lead a model to make unfair or inaccurate predictions
  - Favors one group over others

COMPONENTS OF A MODEL CARD



SAMPLE MODEL CARD #1

Model Card for Census Income Classifier

Model Details

Overview

This is a wide and deep Keras model which aims to classify whether or not an individual has an income of over $50,000 based on various demographic features. The model is trained on the UCI Census Income Dataset. This is not a production model, and this dataset has traditionally only been used for research purposes. In this Model Card, you can review quantitative components of the model's performance and data, as well as information about the model's intended uses, limitations, and ethical considerations.

Version

name: 36dea2e860670aa74691b5695587afe7

Owners

- Model Cards Team, model-cards@google.com

References

- interactive-2020-07-28T20_17_47.911887

Considerations

Use Cases

- This dataset that this model was trained on was originally created to support the machine learning community in conducting empirical analysis of ML algorithms. The Adult Data Set can be used in fairness-related studies that compare inequalities across sex and race, based on people's annual incomes.

Limitations

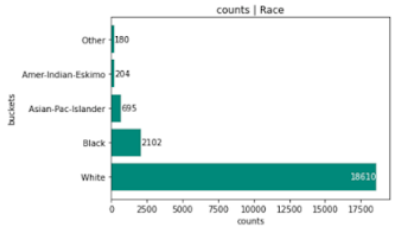
- This is a class-imbalanced dataset across a variety of sensitive classes. The ratio of male-to-female examples is about 2:1 and there are far more examples with the "white" attribute than every other race combined. Furthermore, the ratio of $50,000 or less earners to $50,000 or more earners is just over 3:1. Due to the imbalance across income levels, we can see that our true negative rate seems quite high, while our true positive rate seems quite low. This is true to an even greater degree when we only look at the "female" sub-group, because there are even fewer female examples in the $50,000+ earner group, causing our model to overfit these examples. To avoid this, we can try various remediation strategies in future iterations (e.g. undersampling, hyperparameter tuning, etc), but we may not be able to fix all of the fairness issues.

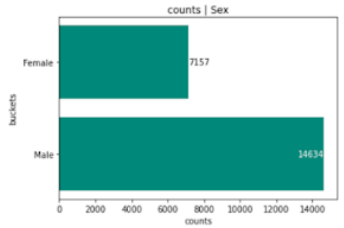
Ethical Considerations

- Risk: We risk expressing the viewpoint that the attributes in this dataset are the only ones that are predictive of someone's income, even though we know this is not the case.
  Mitigation Strategy: As mentioned, some interventions may need to be performed to address the class imbalances in the dataset.

Train Set

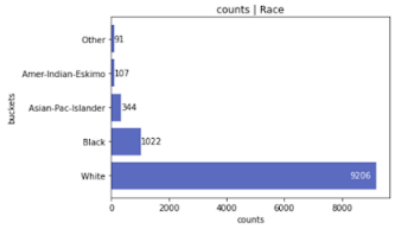
This section includes graphs displaying the class distribution for the "Race" and "Sex" attributes in our training dataset. We chose to show these graphs in particular because we felt it was important that users see the class imbalance.

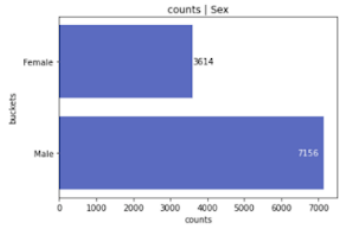




Eval Set

Like the training set, we provide graphs showing the class distribution of the data we used to evaluate our model's performance.





SAMPLE MODEL CARD #2

Object Detection

Model Card v0 Cloud Vision API

- Overview
- Limitations
- Performance
- Test your own images
- Provide feedback

Explore

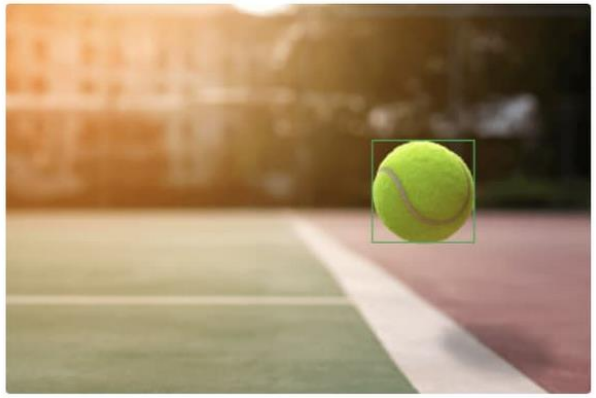
- Face Detection
- About Model Cards

Object Detection

The model analyzed in this card detects one or more physical objects within an image, from apparel and animals to tools and vehicles, and returns a box around each object, as well as a label and description for each object.

On this page, you can learn more about how the model performs on different classes of objects, and what kinds of images you should expect the model to perform well or poorly on.

MODEL DESCRIPTION



Input: Photo(s) or video(s)

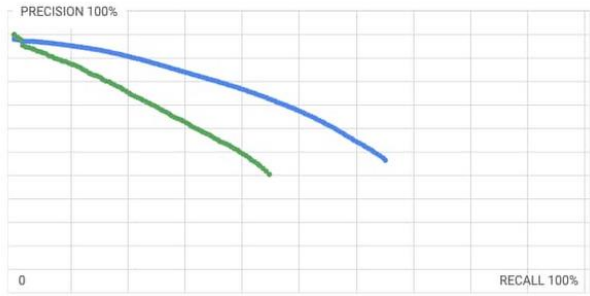
Output: The model can detect 550+ different object classes. For each object detected in a photo or video, the model outputs:

- Object bounding box coordinates
- Knowledge graph ID ("MID")
- Label description
- Confidence score

Model architecture: Single shot detector model with a Resnet 101 backbone and a feature pyramid network feature map.

View public API documentation

PERFORMANCE



Open Images Google Internal

Performance evaluated for specific object classes recognized by the model (e.g. shirt, muffin), and for categories of objects (e.g. apparel, food).

Two performance metrics are reported:

- Average Precision (AP)
- Recall at 60% Precision

Performance evaluated on two datasets distinct from the training set:

- Open Images Validation set, which contains ~40k images and 600 object classes, of which the model can recognize 518.
- An internal Google dataset of ~5,000 images of consumer products, containing 210 object classes, all of which model can recognize.

Go to performance

SAMPLE MODEL CARD #3

Model Details

Overview

This model predicts whether breast cancer is benign or malignant based on image measurements.

Version

name: 95a5ea66-75d5-49d2-af98-17395ea3d7e1
date: 2021-10-21

Owners

Model Cards Team, model-cards-team@email.com

References

- https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)
- https://minds.wisconsin.edu/bitstream/handle/1793/59692/TR1131.pdf

Considerations

Intended Users

- Medical professionals
- ML researchers

Use Cases

- Breast cancer diagnosis

Limitations

- Breast cancer diagnosis

Ethical Considerations

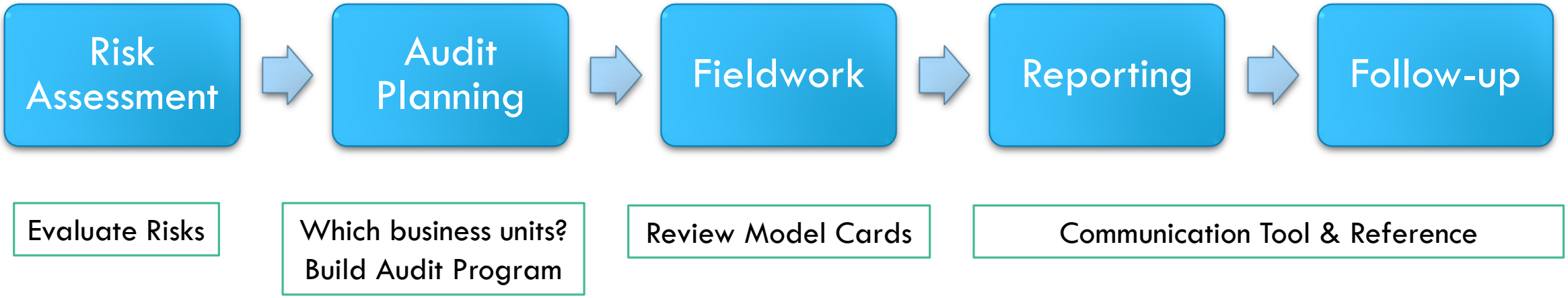
- **Risk:** Manual selection of image sections to digitize could create selection bias
  Mitigation Strategy: Automate the selection process

MODEL CARD USAGE

- Understand Model Behavior
- Evaluate risks/biases and validate compliance
- Review and standardize model card template
- Use as a communication tool during audits

INTEGRATING WITH THE AUDIT PROCESS

Which stage?



KEY TAKEAWAYS

- Model cards improve transparency
  - Easier to digest
  - Quicker to understand
  - Gain insights into behaviour and biases

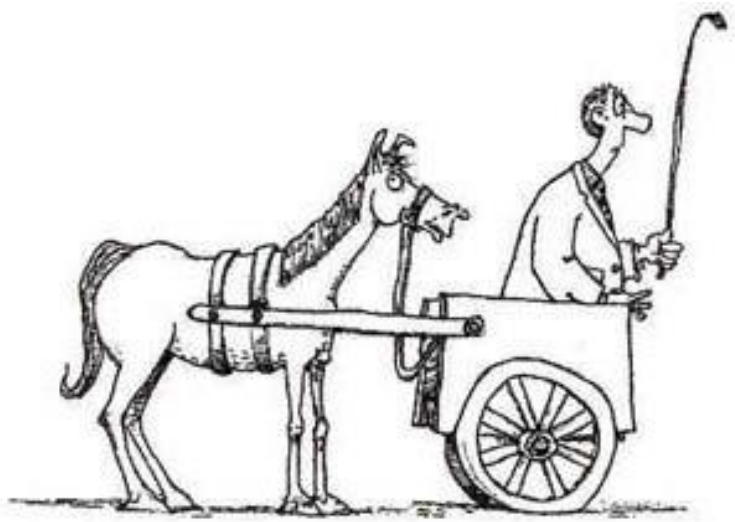
- Promotes value-centric auditing practices

POLLING QUESTION #3

Do you have an AI Governance program in place?

- ❑ Yes
- ❑ No
- ❑ Don’t know

AND REMEMBER…



MODEL CARD ELEMENTS DEFINITION

Name & Type: Identifies the AI model and its category, aiding in classification and differentiation among various models.

Purpose & Description: Outlines the intended use and functionality of the model, providing context for its application.

Data: Describes the dataset(s) used to train and evaluate the model, including sources, size, and characteristics.

Performance: Provides metrics and benchmarks to assess the model's effectiveness and accuracy in various tasks.

Biases: Identifies potential biases present in the model's training data or algorithms, highlighting areas for scrutiny and mitigation.

Limitations: Acknowledges the constraints and constraints of the model, such as its scope, applicability, and potential shortcomings.

Updates & Maintenance: Details plans and procedures for updating and maintaining the model over time, ensuring its relevance and reliability.