KM
UTT

cpe

Computer Engineering

SINCE 1986

Department of Computer Engineering

Machine Learning Operations

Instructors

Asst. Prof. Dr. Santitham Prom-on

santitham.pro@kmutt.ac.th

Dr. Aye Hninn Khine

aye.hnin@kmutt.ac.th

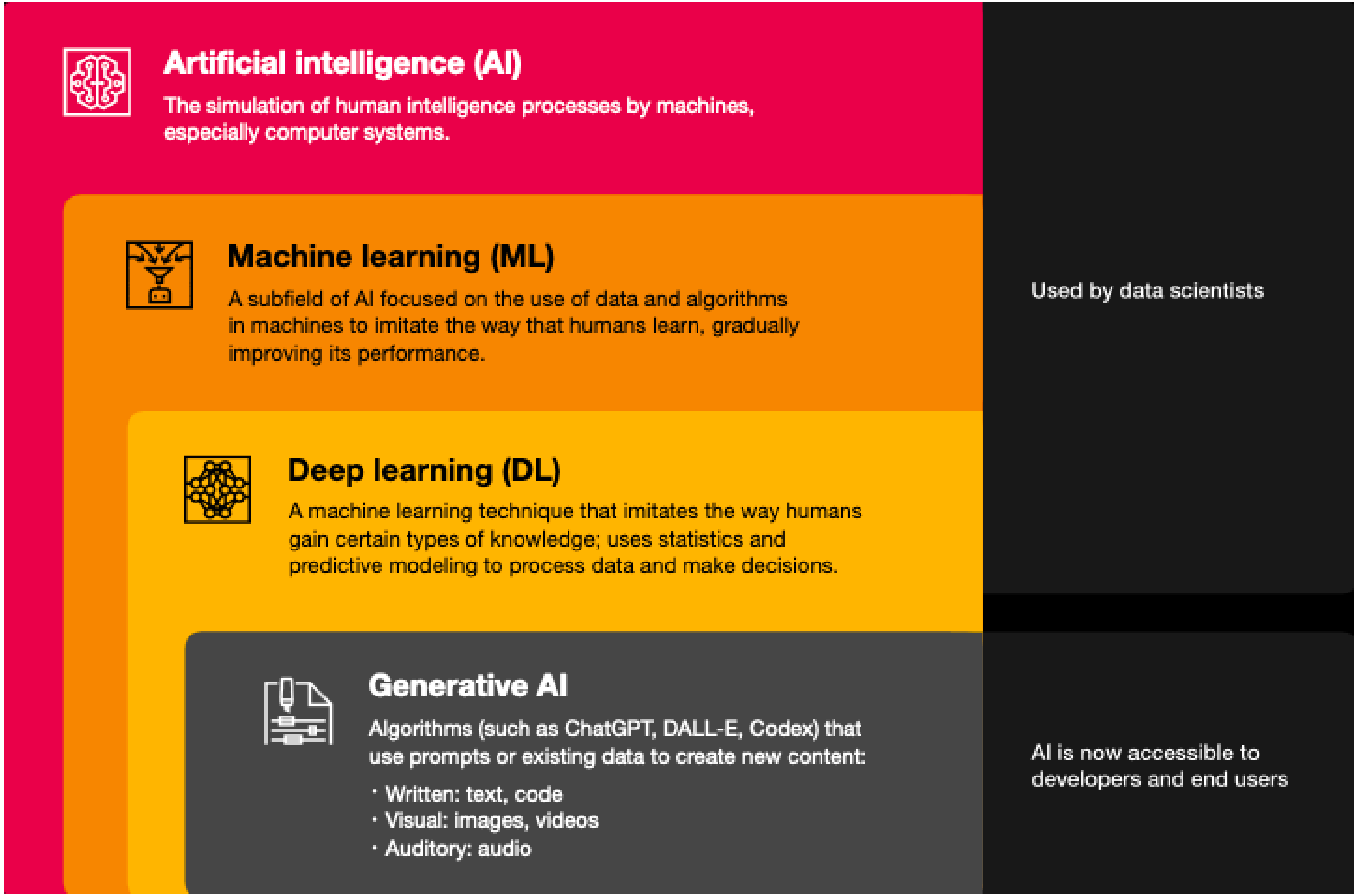
Introduction to Machine Learning System

- Machine Learning Project Life Cycle
- When to use ML?
- Different roles in ML Team
- ML in Research VS ML in Production
- ML VS Traditional Software
- Business Metrics and ML Metrics
- How to frame an ML problem
- Key Features of MLOps

Week 2

Introduction to Machine Learning System

Machine Learning



Artificial intelligence (AI)

The simulation of human intelligence processes by machines, especially computer systems.

Machine learning (ML)

A subfield of AI focused on the use of data and algorithms in machines to imitate the way that humans learn, gradually improving its performance.

Deep learning (DL)

A machine learning technique that imitates the way humans gain certain types of knowledge; uses statistics and predictive modeling to process data and make decisions.

Generative AI

Algorithms (such as ChatGPT, DALL-E, Codex) that use prompts or existing data to create new content:

- Written: text, code
- Visual: images, videos
- Auditory: audio

Used by data scientists

AI is now accessible to developers and end users

Source - https://www.pwc.com/us/en/services/governance-insights-center/library/assets/pwc-the-power-of-ai-and-gen-ai-what-boards-should-know.pdf

Machine Learning

- ML is an approach to (1) **learn** (2) **complex patterns** from (3) **existing data** and use the patterns to make (4) **predictions** on (5) **unseen data**.

When to use ML?

- Learn: the system has the capacity to learn
- Complex: there are patterns to learn, and they are complex
- Existing data: data is available, or it’s possible to collect the data
- Predictions: It is a prediction problem
- Unseen data: unseen data shares patterns with the training data
- It’s repetitive
- The cost of wrong prediction is cheap
- It’s at scale
- The patterns are constantly changing

When not to use ML?

- It’s unethical.
- Simpler solutions do the trick.
- It’s not cost effective.

ML in the class VS ML in the industry

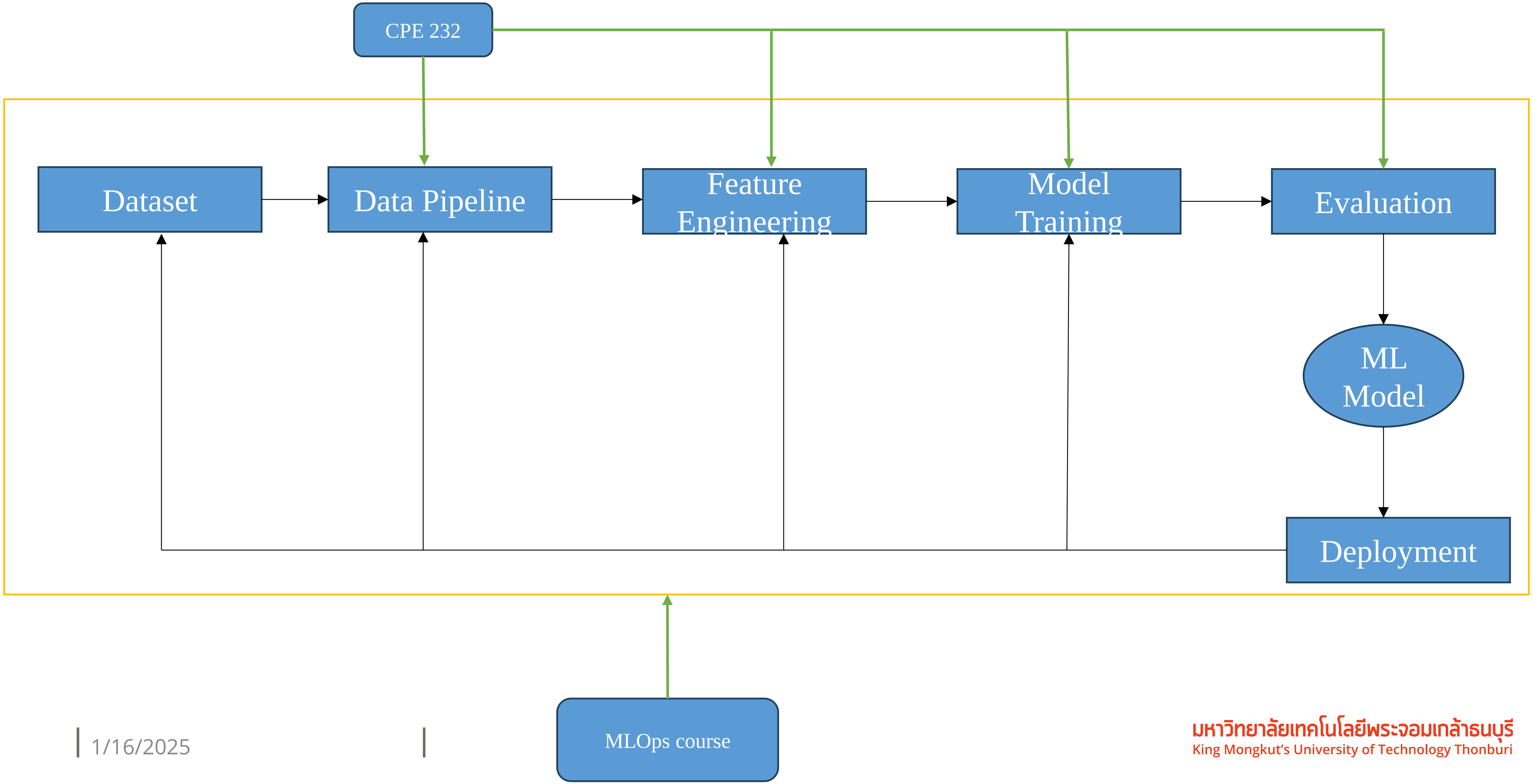
ML use cases

- Recommendation Engine/Recommender System
- Predictive Typing
- Machine Translation
- Fraud Detection
- Smart Health

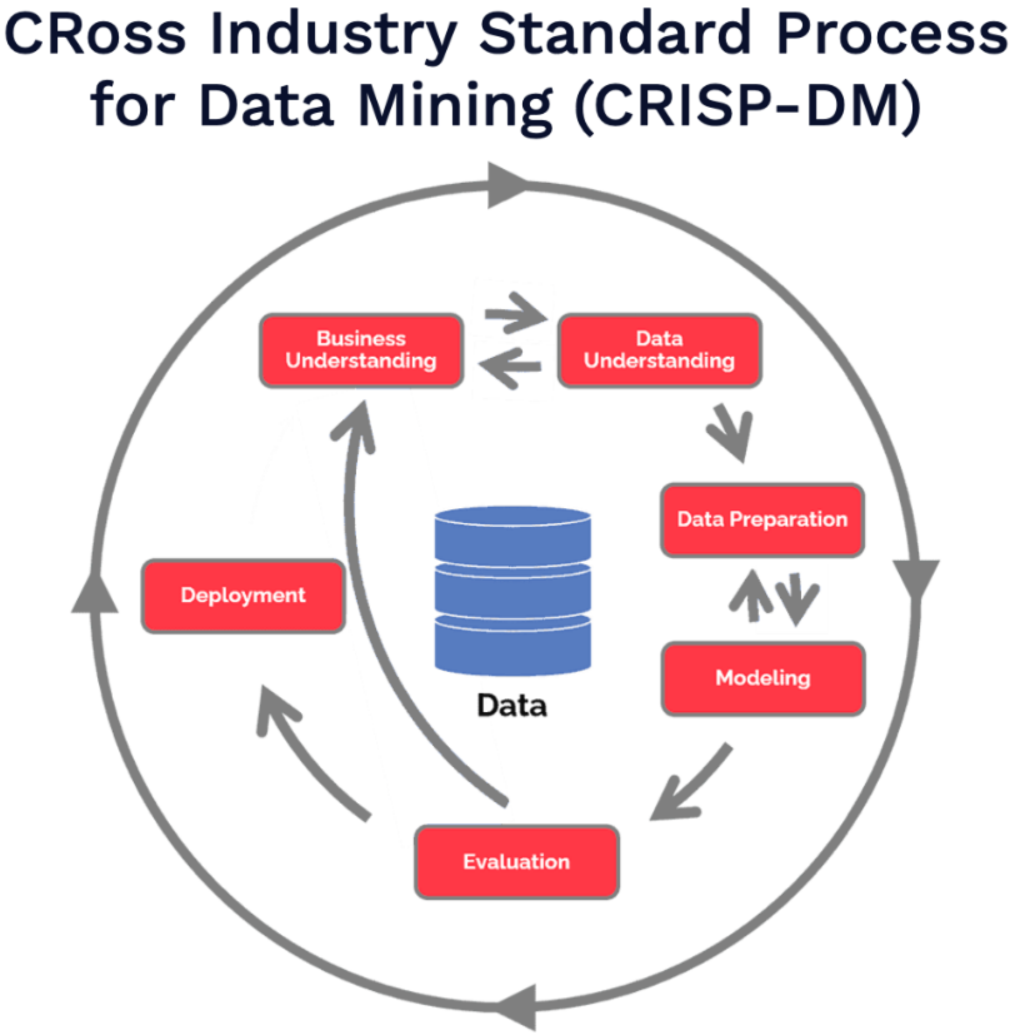
ML System Requirement

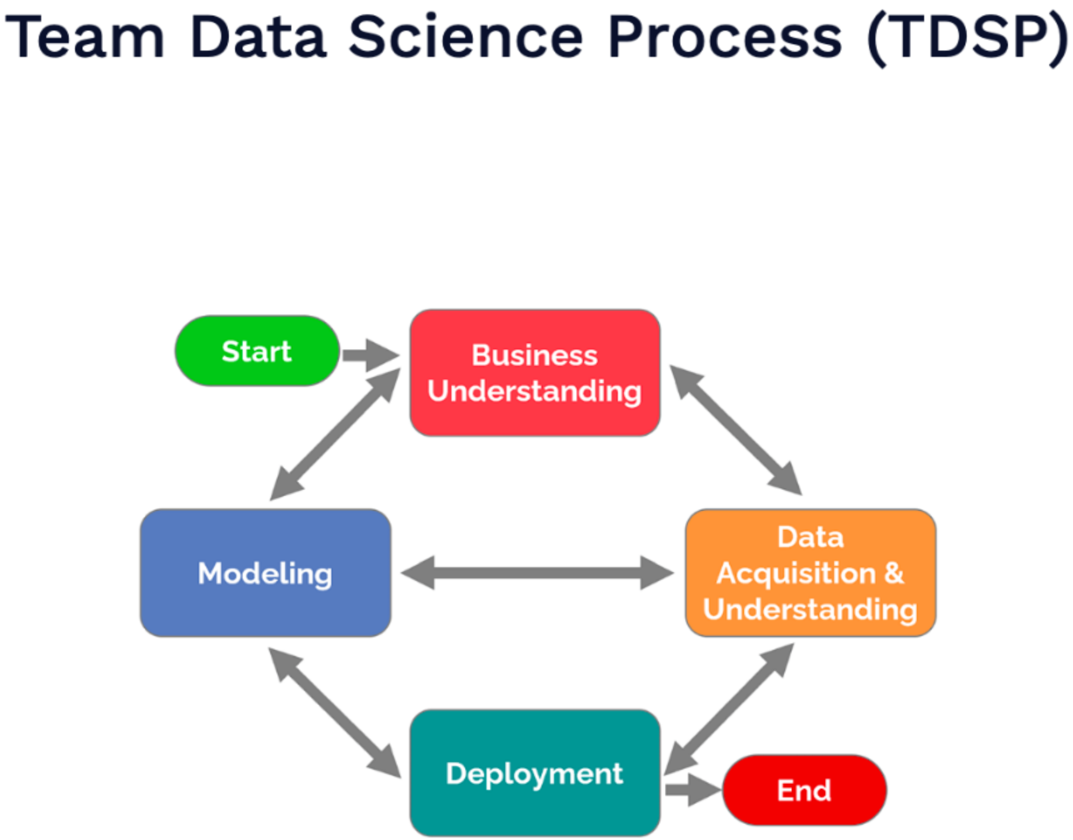
- Reliability – The system should continue to perform the correct function at the desired level of performance even in the face of adversity
- Scalability – resource scaling (up scale/down scale)
- Maintainability – set up your infra that different contributors can work using tools that they are comfortable with
- Adaptability – allowing updates without service interruption

ML Project Life Cycle



Agile for ML




Team Data Science Process (TDSP)
Start
Business Understanding
Data Acquisition & Understanding
Modeling
Deployment
End

https://fullstackdeeplearning.com/course/2022/lecture-8-teams-and-pm/#how-to-manage-ml-teams-better

Iterative ML development

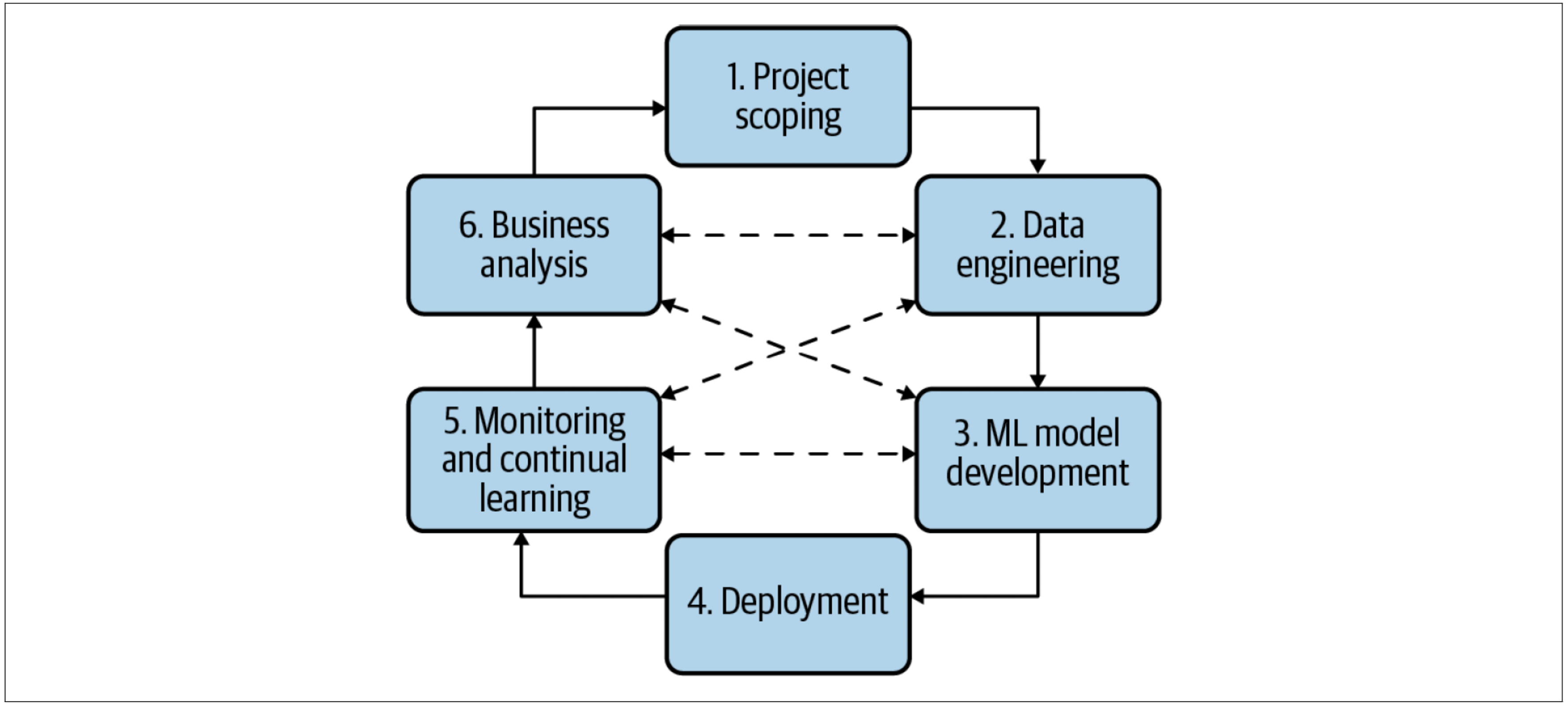
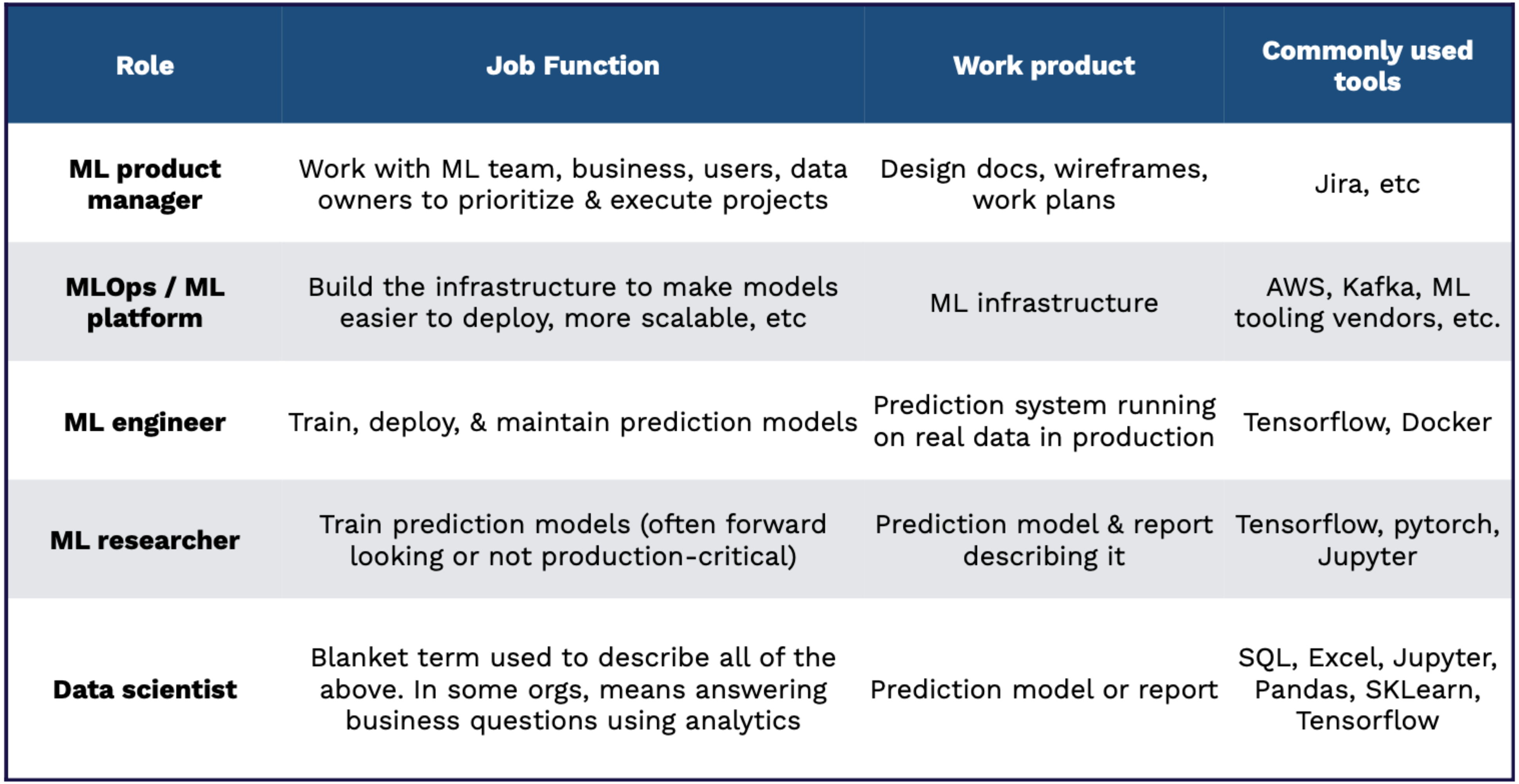


Figure 2-2. The process of developing an ML system looks more like a cycle with a lot of back and forth between steps

ML Team



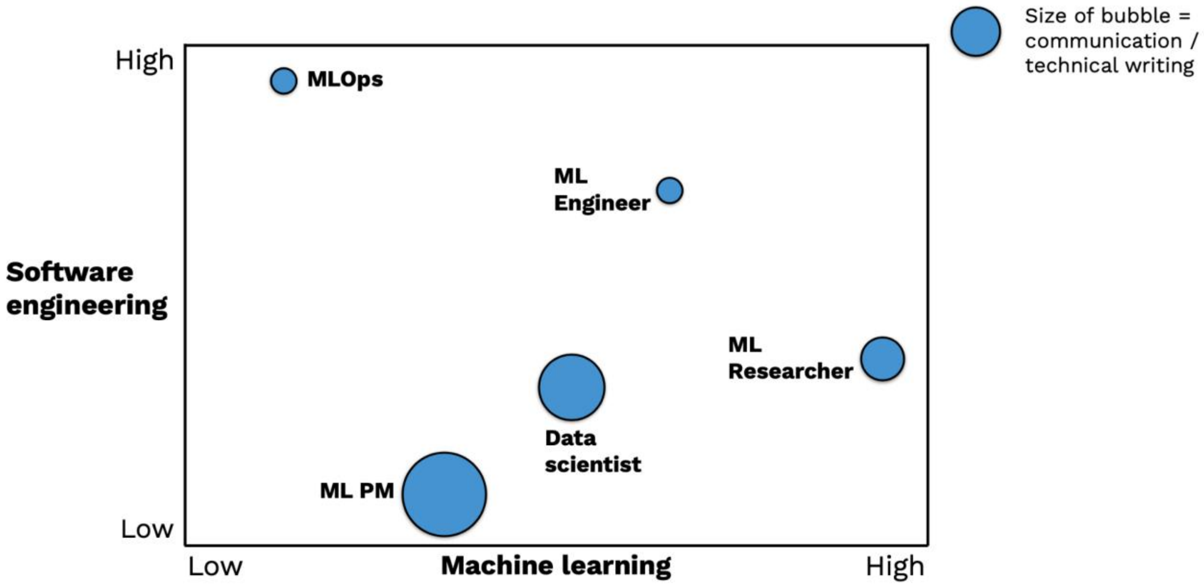
Role	Job Function	Work product	Commonly used tools
ML product manager	Work with ML team, business, users, data owners to prioritize & execute projects	Design docs, wireframes, work plans	Jira, etc
MLOps / ML platform	Build the infrastructure to make models easier to deploy, more scalable, etc	ML infrastructure	AWS, Kafka, ML tooling vendors, etc.
ML engineer	Train, deploy, & maintain prediction models	Prediction system running on real data in production	Tensorflow, Docker
ML researcher	Train prediction models (often forward looking or not production-critical)	Prediction model & report describing it	Tensorflow, pytorch, Jupyter
Data scientist	Blanket term used to describe all of the above. In some orgs, means answering business questions using analytics	Prediction model or report	SQL, Excel, Jupyter, Pandas, SKLearn, Tensorflow

https://fullstackdeeplearning.com/course/2022/lecture-8-teams-and-pm/

People of MLOps

- Subject Matter Experts – Provide business questions, goals or KPI around which ML models should be framed.
- Data Scientists/ML researchers – Build models that address the business question or needs brought by SMEs. Assess model quality in tandem with SMEs to ensure they answer initial business questions or needs.
- Data Engineers – Build ETL (Extract, Transform, Load) ETL pipeline and optimize the retrieval and use of data to power ML models.
- ML Engineers (Software Engineers with ML Skills) – Conduct and build operational systems and test for security, performance, availability. CI/CD pipeline management. Ensure a scalable and flexible environment for ML model pipelines, from design to development and monitoring.

Skills Required



- The **MLOps** is primarily a software engineering role, which often comes from a standard software engineering pipeline.
- The **ML Engineer** requires a rare mix of ML and Software Engineering skills. This person is either an engineer with significant self-teaching OR a science/engineering Ph.D. who works as a traditional software engineer after graduate school.
- The **ML Researcher** is an ML expert who usually has an MS or Ph.D. degree in Computer Science or Statistics or finishes an industrial fellowship program.
- The **ML Product Manager** is just like a traditional Product Manager but with a deep knowledge of the ML development process and mindset.
- The **Data Scientist** role constitutes a wide range of backgrounds, from undergraduate to Ph.D. students.

https://fullstackdeeplearning.com/course/2022/lecture-8-teams-and-pm

ML Systems VS Traditional Software

- Traditional software → code and data are separated
- ML systems → part data, part code
  - Data dependencies cost more than code dependencies
  - Unstable data dependencies – data are changing over time
  - ML system is a software 2.0 (Data + Code)

Key differences between ML in research and ML in production

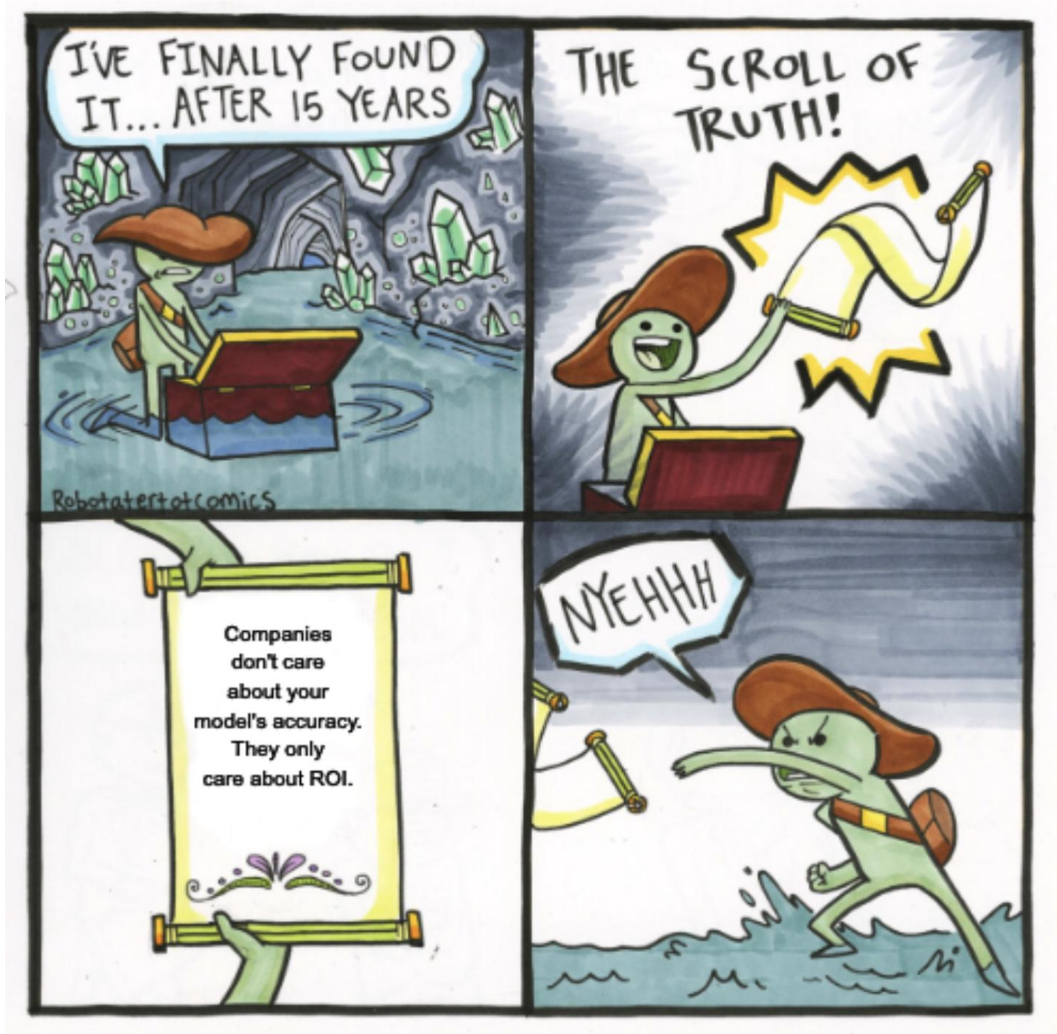
	Research	**Production**
Requirements	Model performance	Different stakeholders have different objectives
Computational priority	Fast training, high throughput	Fast inference, low latency
Data	Static	Constantly shifting
Fairness	Often not a focus	Must be considered
Interpretability	Often not a focus	Must be considered

Different Metrics in ML

- Different stakeholders → different requirements
- Example - Food ordering app
- ML engineers – want a model that recommends restaurants that users will most likely order from, and believe they can do so by using a more complex model with more data.
- Sales team – wants a model that recommends the more expensive restaurants since these restaurants bring in more service fees.
- Product team – wants a model that return the recommended restaurants in less than 100 milliseconds.
- ML platform team – want to hold off on model updates to prioritize improving the ML platform.
- Manager – wants to maximize the margin, and one way to achieve this might be to let go of the ML team

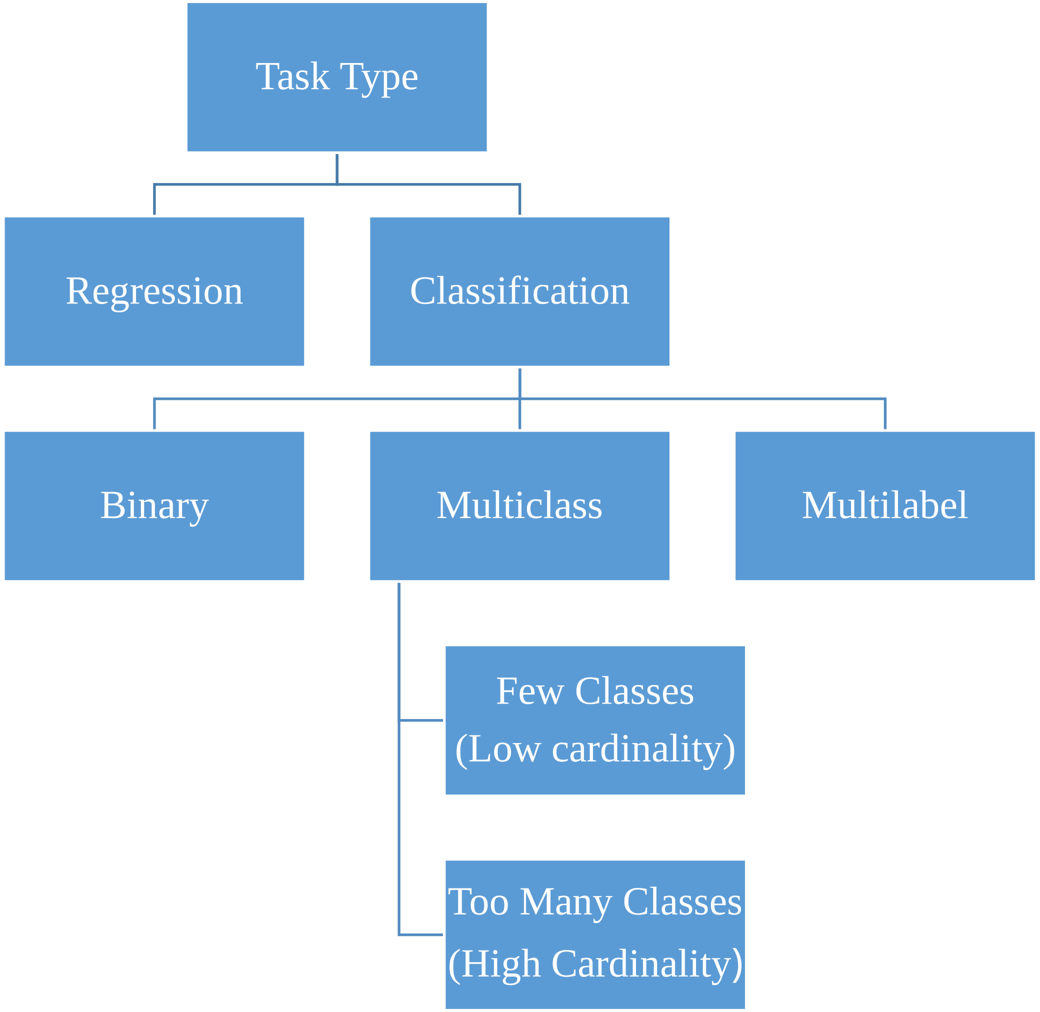
Business and ML objectives

- ML objective
  - Accuracy, F1
  - Latency
- Business objective
  - To maximize profits of shareholders
  - Increase sales
  - Cutting costs
  - High customer satisfaction
  - Increase time spent


I'VE FINALLY FOUND IT... AFTER 15 YEARS
THE SCROLL OF TRUTH!
Companies don't care about your model's accuracy. They only care about ROI.
NYEHHH
Robotatertotcomics

Framing Machine Learning Problem

Type of ML Tasks


Task Type
Regression
Classification
Binary
Multiclass
Multilabel
Few Classes (Low cardinality)
Too Many Classes (High Cardinality)

Bank Customer Satisfaction

- Problem
  - Slow customer support
  - Competitor bank process their customer requests two times faster
  - Routing customer requests to the right department among 4 departments: accounting, inventory, HR, and IT
- ML solution
  - Develop an ML model to predict which of these four departments as a request should go to
  - Output – department the request should go to
  - Classification Problem

My painful experiences

- Problem – Hate Speech Detection from Social Media (Burmese, Sinhala, and Tamil languages)
- ML Solution – Classification Problem (Binary – Hate speech/Non-Hate Speech) / Regression (A numerical score for hate speech intensity)
- Challenges
  - How to define hate speech?? (Took more than 4 months to get data labelling guideline)
  - Binary classification is not enough
  - Not all countries have the same hate speech scenarios (need political context knowledge)
  - hierarchical classification or multiclass classification are requested by the clients
  - A new category will be added in the future which leads to re-training the model from the beginning

Handling High Cardinal Classification Problem

- ML models typically need at least 100 examples for each class to learn to classify that class
- Example – Product classification (100 classes) → 100x100 = 10000 training examples
- Hierarchical classification (Fashion -> Women -> Blouse)
- Multiple Binary classification models (Fashion? Electronic Gadgets? Kitchen Utilities?)

Multiple ways to frame a problem

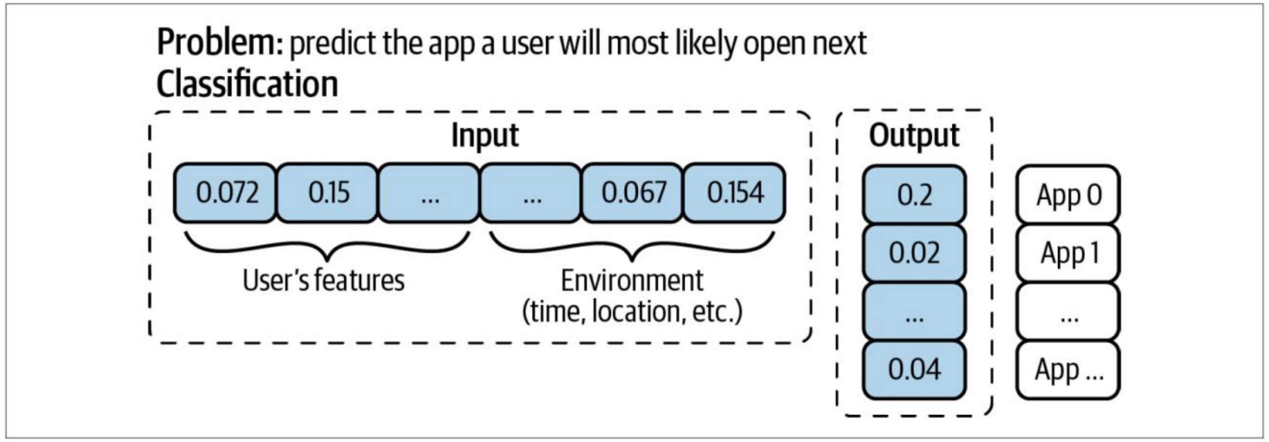


Figure 2-5. Given the problem of predicting the app a user will most likely open next, you can frame it as a classification problem. The input is the user's features and environment's features. The output is a distribution over all apps on the phone.

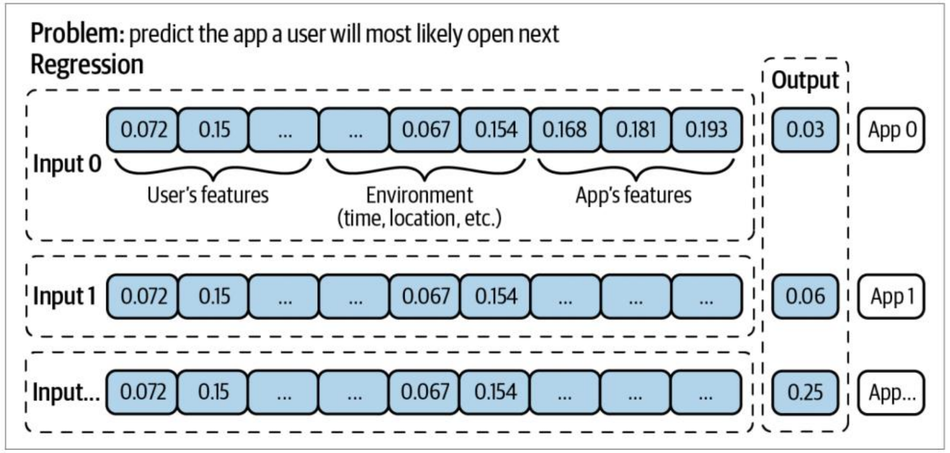


Figure 2-6. Given the problem of predicting the app a user will most likely open next, you can frame it as a regression problem. The input is the user's features, environment's features, and an app's features. The output is a single value between 0 and 1 denoting how likely the user will be to open the app given the context.

Key Features of MLOps

MLOps

- Standardization and streamlining of ML life cycle management
- Data scientists and machine learning engineers can collaborate and increase the pace of model development and production, by implementing continuous integration and deployment (CI/CD) practices with proper monitoring, validation, and governance of ML models.

MLOps is a culture

Why MLOps?

- Risk mitigation
  - The risk that the model is unavailable for a given period of time
  - The risk that the model returns a bad prediction for a given sample
  - The risk that the model accuracy or fairness decreases over time
  - The risk that the skills necessary to maintain the model (e.g., data science talent) are lost
- Keep track of versioning, especially with experiments in the design phase.
- Understand whether retrained models are better than the previous versions
- Ensure that model performance is not degrading in production

Establishing Business Objectives

- Reducing fraudulent transactions to <0.1 %
- Identifying people’s faces on their social media photos
- Business objectives normally comes with performance targes, technical infrastructure requirements, and cost constraints

Data Sources and Exploratory Data Analysis

- What relevant datasets are available?
- Is this data sufficiently accurate and reliable?
- How can stakeholders get access to this data?
- What data properties (features) can be made available by combining multiple sources of data?
- Will this data be available in real time?
- Is there a need to label some of the data with the “ground truth” that is to be predicted, or does unsupervised learning make sense? If so, how much will this cost in terms of time and resources?
- What platform should be used?
- How will data be updated once the model is deployed?
- Will the use of the model itself reduce the representativeness of the data?
- How will the KPIs, which are established along with the business objectives, be measured?

Feature Engineering and Selection

- EDA naturally leads into feature engineering and feature selection.
- Process of taking taw data from the selected datasets and transforming it into “features” that better represent the underlying problem to be solved.

Training and Evaluation

- Several algorithms may be tested
- Features can be automatically generated and hyperparameters tuned
- Keep track of the results of each experiment (Experiment tracking)
- An experiment tracking tool can simplify the process of remembering the data, the features selection, and model parameters alongside the performance metrics.
- Save enough information about the environment of the model was developed in so that the model can be reproduced with the same results from scratch (Reproducibility)
- Version control of all the data and parameters involved, including the data used to train and evaluate the model, as well as a record of the software environment.
- Partial dependence plots which look at the impact of features on the predicted outcome (Explainability)
- Individual model predictions, such as Shapley values, which explain how the value of each feature contributes to a specific prediction
- What-if analysis, which helps the ML model user to understand the sensitivity of the prediction to its inputs

Productionalization and Deployment

- Domain of SWE and DevOps team
- Information exchange between the data scientists and these teams must not be underestimated.
- Model-as-a-service, or live-scoring model
  - Typically the model is deployed into a simple framework to provide a REST API endpoint that responds to requests in real time.
- Embedded Model
  - Here the model is packaged into an application, which is then published. A common example is an application that provides batch-scoring of requests
- Containerization technologies such as Docker are light-weight alternatives to virtual machines, allowing applications to be deployed in independent, self-contained environments, matching the exact requirements of each model.

Monitoring (DevOps)

- Is the model getting the job done quicky enough?
- Is it using a sensible amount of memory and processing time?

Monitoring (Data Scientist)

- ML models can degrade over time, since ML models are effectively models of the data they were trained on.
- If the training data reflects the real world well, then the model should be accurate, and useful.
- The training data used to build a fraud detection model six months ago won’t reflect a new type of fraud that has started to occur in the last three months.
- In many use cases, such as fraudulent transaction, obtaining ground truth is much slower.
- Identifying input drift is one of the most important components of an adaptable MLOps strategy.

Monitoring (Business)

- Is the model delivering value to the enterprise?
- Do the benefits of the model outweigh the cost of developing and deploying it?

The Feedback Loop

- Shadow testing
  - The new model is deployed into the live environment alongside the existing model.
  - Each new request is scored by the new model and the results logged, but not returned to the requestor.
  - The results of existing model and new model can be compared statistically.
- A/B Testing
  - Both models are deployed into the live environment
  - Input requests are split between the two models.

Governance

- Set of controls placed on a business to ensure that it delivers on its responsibilities to all stakeholders, from shareholders and employees to the public and national governments.
- Personal data protection laws
  - General Data Protection Law (GDPR) for EU countries
  - Personal Data Protection Act (PDPA) in Thailand
- Governance Initiatives in MLOps
  - Data Governance
    - A framework for ensuring appropriate use and management of data
  - Process Governance
    - The use of well-defined processes to ensure all governance considerations have been addressed at the correct point in the life cycle of the model and that a full and accurate record has been kept

Data Governance

- Can the selected datasets be used for ML development?
- What are the terms of use?
- Is there personally identifiable information (PII) that must be redacted or anonymized?
- Are there features, such as gender, that legally cannot be used in this business context?
- Are minority populations sufficiently well represented that the model has equivalent performances on each group?

ML recruitment model famously discriminated against women by identifying all female schools which were less represented in the company's upper management which reflected the historical dominance of men in the organization. (Reuter News)

Process Governance

- Formalizing the steps in MLOps process and associating actions with them.
- To ensure every governance-related consideration is made at the correct time and correctly acted upon.
- To enable oversight from outside of the strict MLOps process.

Human side of Machine Learning

User experience

- Consistency
  - Specify the conditions in which the system must return the same predictions
  - Returns the frequently ordered restaurant for a specific menu in food ordering app
- Human-in-the-loop
  - Let the users evaluate the output and give feedback to it (E.g., Generative AI models such as GPT)
- Smooth failing
  - Having a back up model which is simpler in architecture and faster inference time than the main model
  - Speed—accuracy trade off
  - Faster model might give users worse predictions but might still be preferred om situations where latency is crucial.

Team Structure

- Approach 1: Have a separate team to manage production
  - Communication and coordination overhead – A team can become blockers for other teams
  - Debugging challenges – When something fails, you don’t know whether your team’s code or some other team’s code might have caused it.
  - Finger-pointing– Even when you’ve figured out what went wrong, each team might think it’s another team’s responsibility to fix it.
  - Narrow context – No one has visibility into the entire process to optimize/improve it.
- Approach 2: Data scientists own the entire process (end-to-end data scientists)
  - Not practical
  - In theory, you can learn both sets of skills (model development MLOps)
  - In practice, the more time you spend on one means the less time you spend on the other,

Responsible AI

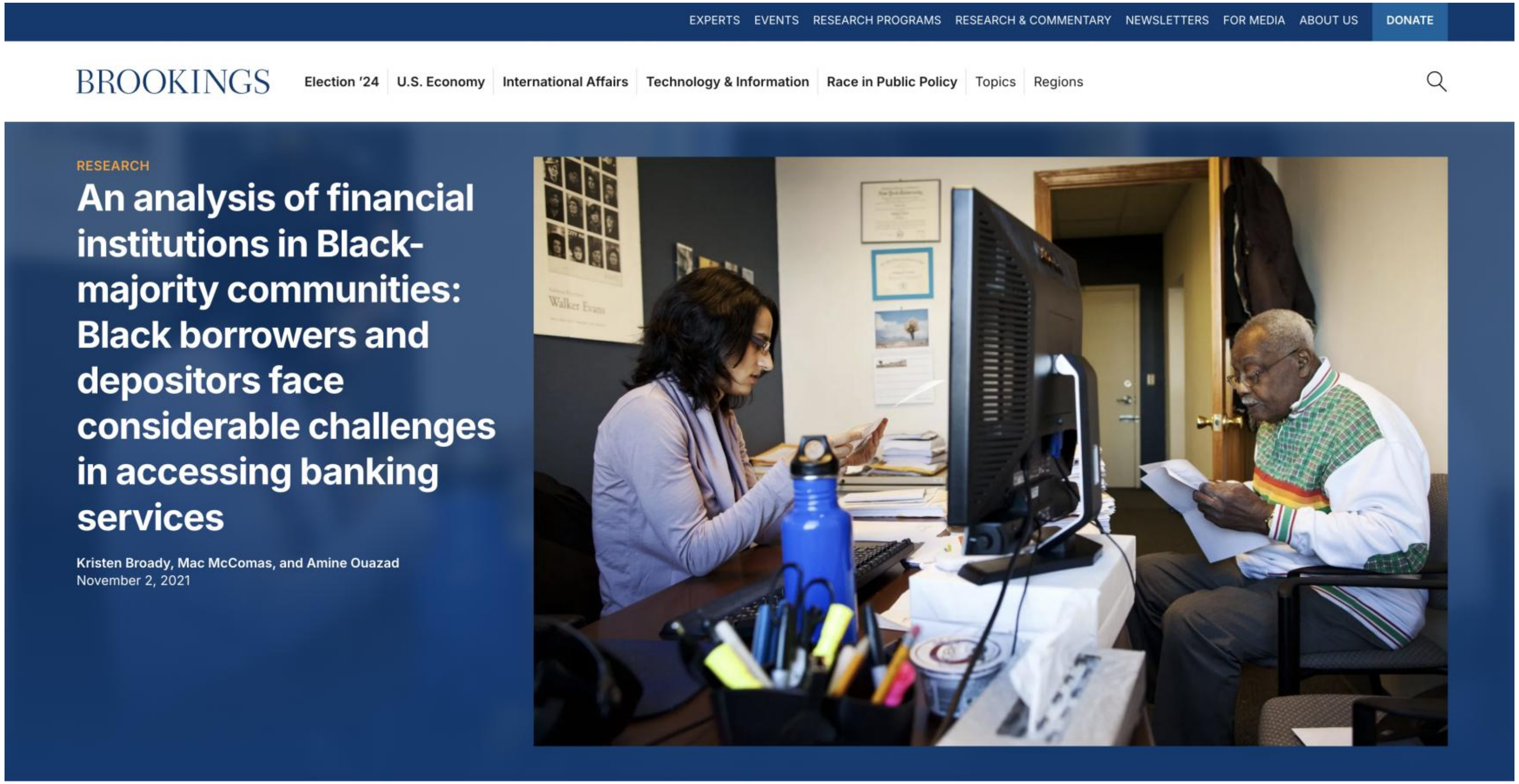
Fairness

Privacy

Transparency

Accountability

AI incident database- https://incidentdatabase.ai/



RESEARCH

An analysis of financial institutions in Black-majority communities: Black borrowers and depositors face considerable challenges in accessing banking services

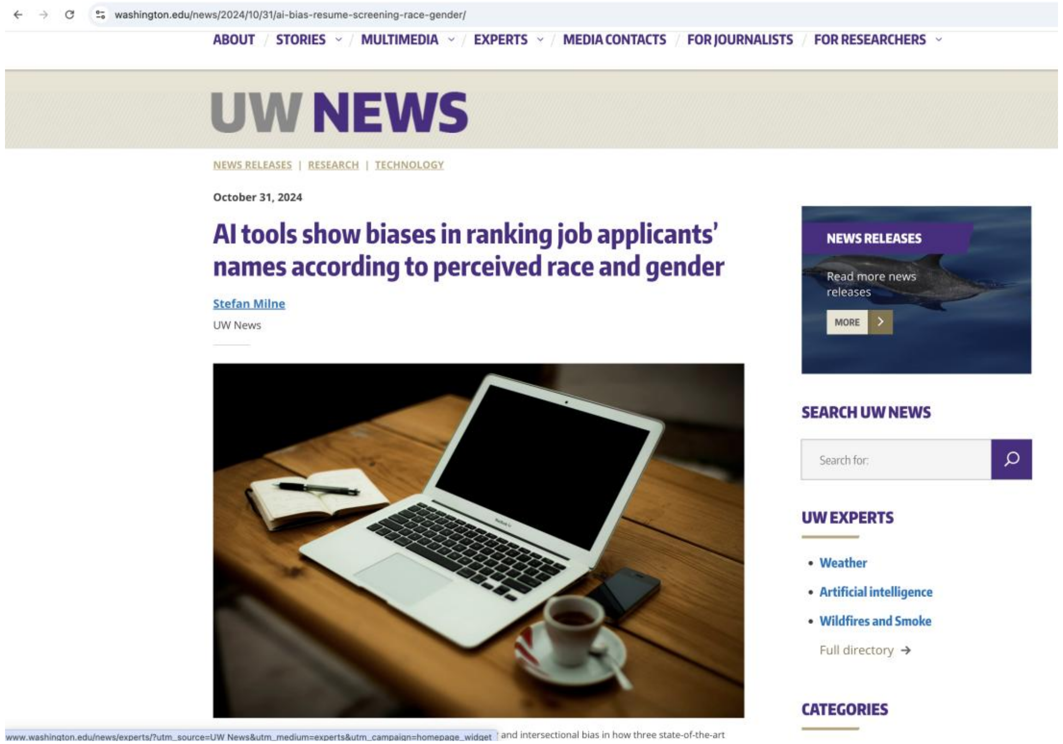
Kristen Broady, Mac McComas, and Amine Ouazad

November 2, 2021

https://www.brookings.edu/articles/an-analysis-of-financial-institutions-in-black-majority-communities-black-borrowers-and-depositors-face-considerable-challenges-in-accessing-banking-services/

Irresponsible AI: Case Studies

AI bias in job applicant’s ranking


washington.edu/news/2024/10/31/ai-bias-resume-screening-race-gender/
ABOUT / STORIES / MULTIMEDIA / EXPERTS / MEDIA CONTACTS / FOR JOURNALISTS / FOR RESEARCHERS
UW NEWS
NEWS RELEASES | RESEARCH | TECHNOLOGY
NEWS RELEASES
Read more news releases
MORE
SEARCH UW NEWS
Search for:
UW EXPERTS
Weather
Artificial intelligence
Wildfires and Smoke
Full directory →
CATEGORIES
www.washington.edu/news/experts/?utm_source=UW News&utm_medium=experts&utm_campaign=homepage_widget

October 31, 2024

AI tools show biases in ranking job applicants’ names according to perceived race and gender

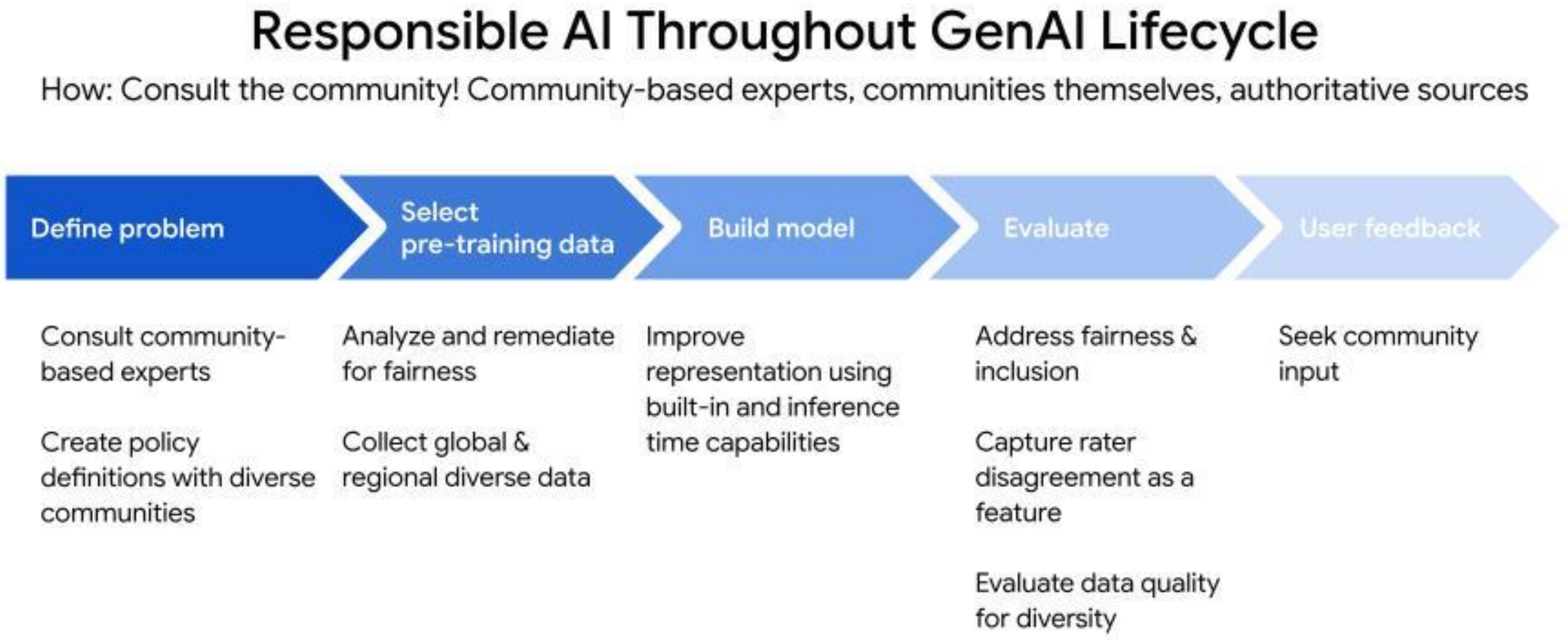
Stefan Milne

UW News

and intersectional bias in how three state-of-the-art

- Automated Grader System Bias (UK-2020)
  - Prioritize school over student performance
  - If school A has historically outperformed school B in the past, Ofqual wanted an algorithm that, on average, also gives students from school A higher grades than students form school B

Responsible AI Framework



Responsible AI Throughout GenAI Lifecycle

How: Consult the community! Community-based experts, communities themselves, authoritative sources

Define problem	Select pre-training data	Build model	Evaluate	User feedback
Consult community-based experts Create policy definitions with diverse communities	Analyze and remediate for fairness Collect global & regional diverse data	Improve representation using built-in and inference time capabilities	Address fairness & inclusion Capture rater disagreement as a feature Evaluate data quality for diversity	Seek community input

Credit: Google AI Research

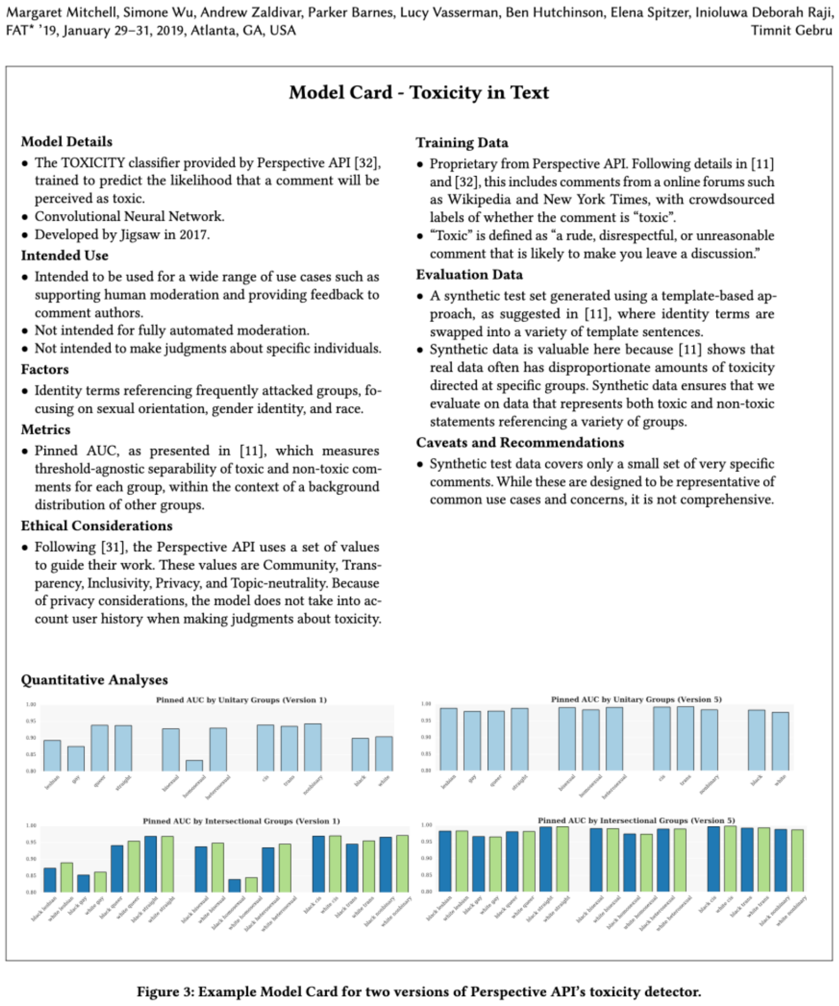
A Framework for Responsible AI

- Discover sources of model biases
  - Training Data – Is the data used for developing your model representative of the data your model will handle in the real world?
  - Labeling – How to measure quality of labels? The more annotators have to rely on their subjective experience, the more room for human biases.
  - Feature Engineering – Does your model use any feature that contains sensitive information? Legally protected classes – ethnicity, gender, religion)
  - Model’s Objective – Are you optimizing your model using an objective that enables fairness to all users?
  - Evaluation – Are you performing adequate, fine-grained evaluation to understand your model’s performance on different groups of users?

A Framework for Responsible AI

- Understand the limitations of the data-driven approach
  - For example, to build an equitable automated grading system, it’s essential to work with domain experts to understand the demographic distribution of the student population and how socioeconomic factors get reflected in the historical performance data
- Act early
  - The earlier in the development cycle of an ML system that you can start thinking about how this system will affect the life of users and what biases your system might have, the cheaper it will be to address these biases.
- Model Card
  - Model cards are short documents accompanying trained ML models that provide information on how these models were trained and evaluated.
  - Model cards also disclose the context in which models are intended to be used.

Example Data Card



Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji,
FAT* '19, January 29–31, 2019, Atlanta, GA, USA
Timnit Gebru

Model Card - Toxicity in Text

Model Details

- The TOXICITY classifier provided by Perspective API [32], trained to predict the likelihood that a comment will be perceived as toxic.
- Convolutional Neural Network.
- Developed by Jigsaw in 2017.

Intended Use

- Intended to be used for a wide range of use cases such as supporting human moderation and providing feedback to comment authors.
- Not intended for fully automated moderation.
- Not intended to make judgments about specific individuals.

Factors

- Identity terms referencing frequently attacked groups, focusing on sexual orientation, gender identity, and race.

Metrics

- Pinned AUC, as presented in [11], which measures threshold-agnostic separability of toxic and non-toxic comments for each group, within the context of a background distribution of other groups.

Ethical Considerations

- Following [31], the Perspective API uses a set of values to guide their work. These values are Community, Transparency, Inclusivity, Privacy, and Topic-neutrality. Because of privacy considerations, the model does not take into account user history when making judgments about toxicity.

Training Data

- Proprietary from Perspective API. Following details in [11] and [32], this includes comments from a online forums such as Wikipedia and New York Times, with crowdsourced labels of whether the comment is "toxic".
- "Toxic" is defined as "a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion."

Evaluation Data

- A synthetic test set generated using a template-based approach, as suggested in [11], where identity terms are swapped into a variety of template sentences.
- Synthetic data is valuable here because [11] shows that real data often has disproportionate amounts of toxicity directed at specific groups. Synthetic data ensures that we evaluate on data that represents both toxic and non-toxic statements referencing a variety of groups.

Caveats and Recommendations

- Synthetic test data covers only a small set of very specific comments. While these are designed to be representative of common use cases and concerns, it is not comprehensive.

Quantitative Analyses

Pinned AUC by Unitary Groups (Version 1)

Pinned AUC by Unitary Groups (Version 5)

Pinned AUC by Intersectional Groups (Version 1)

Pinned AUC by Intersectional Groups (Version 5)

Figure 3: Example Model Card for two versions of Perspective API's toxicity detector.

Reading Assignments

- https://proceedings.neurips.cc/paper_files/paper/2015/file/86df7dcfd896fcaf2674f757a2463eba-Paper.pdf
- https://blog.acolyer.org/2019/10/07/150-successful-machine-learning-models

Discussion (five students per group)

Consider you are a project leader for developing an automated COVID-19 screening system using medical images.

(1) Who will be SMEs? (Subject Matter Expert)

(2) What is your business metrics and objectives?

(3) How will you frame your machine learning model?

(4) What are your machine learning metrics?

(5) How will you collect the data? Is it labeled or unlabeled?

(6) If it is not labeled, how will you get the ground truth data?

(7) What kind of people will you hire for your team?

(8) How will you integrate responsible AI practices to your system?