Lecture 8: Pitfalls and Considerations When Applying ML

CMSC 25910

Spring 2022

The University of Chicago

Outline

- Feature Engineering
- Training data
- Metrics
- Commoditization of ML
- ML in the Wild
    - Concept Drift
    - ML in Pipelines
- Algorithmic Decision Making
    - ML as an artifact within {technical, sociotechnical} systems

Feature Engineering

The Ingredients of an ML application

- (Possibly labeled) Training dataset: [X_i, y_i]
- Model
- Task/Metric, Optimizer

Feature Representation

- Table data -> matrix
- Transform categorical variables into a numerical representation
    - Dummy coding
- Normalization
- Standardization
- Binning
- Other transformations

Same Running Example As Last Time

Name	Age	Department	Gender	Title	**Salary**
Jack	55	CS	M	Professor	??
Jane	27	Stats	F	Assistant Professor	??

x_1, x_2, x_3, x_4, x_5

$\hat{y}$

Variables/attributes/columns become ‘features’ of the input vector

Feature Engineering

- Feature engineering
- Goal: Select the variables to feed into the model
    - More variables do not always lead to better models!

Augmenting Features

- Augment initial data with more features
  - By joining with other datasets

Gender	Title	**Salary**
M	Professor	??
F	Assistant Professor	??

JOIN

Name	Age	Department
Jack	55	CS
Jane	27	Stats

What Features Are You Selecting?

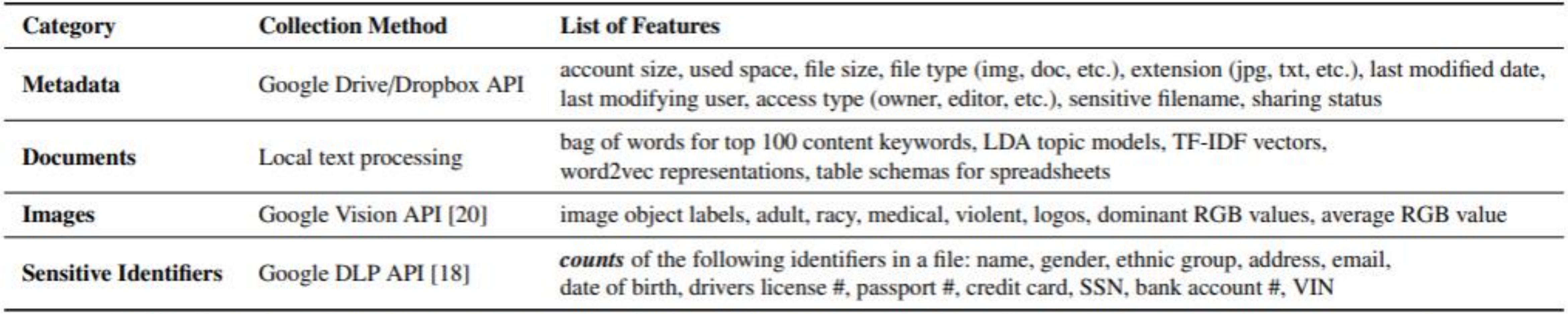
- How can you be sure that sensitive information is not represented in the model?
  - Is removing protected classes enough?
  - Think about information leakage!
- Anonymizing PII
  - Does this solve the problem?
  - Can you anonymize data?

Name	Age	Department	Gender	Title	**Salary**
Jack	55	CS	M	Professor	??
Jane	27	Stats	F	Assistant Professor	??

Feature Engineering

- Preprocessing data
- What aspects of data matter?
  - What aspects **should** matter?

Example of Feature Selection

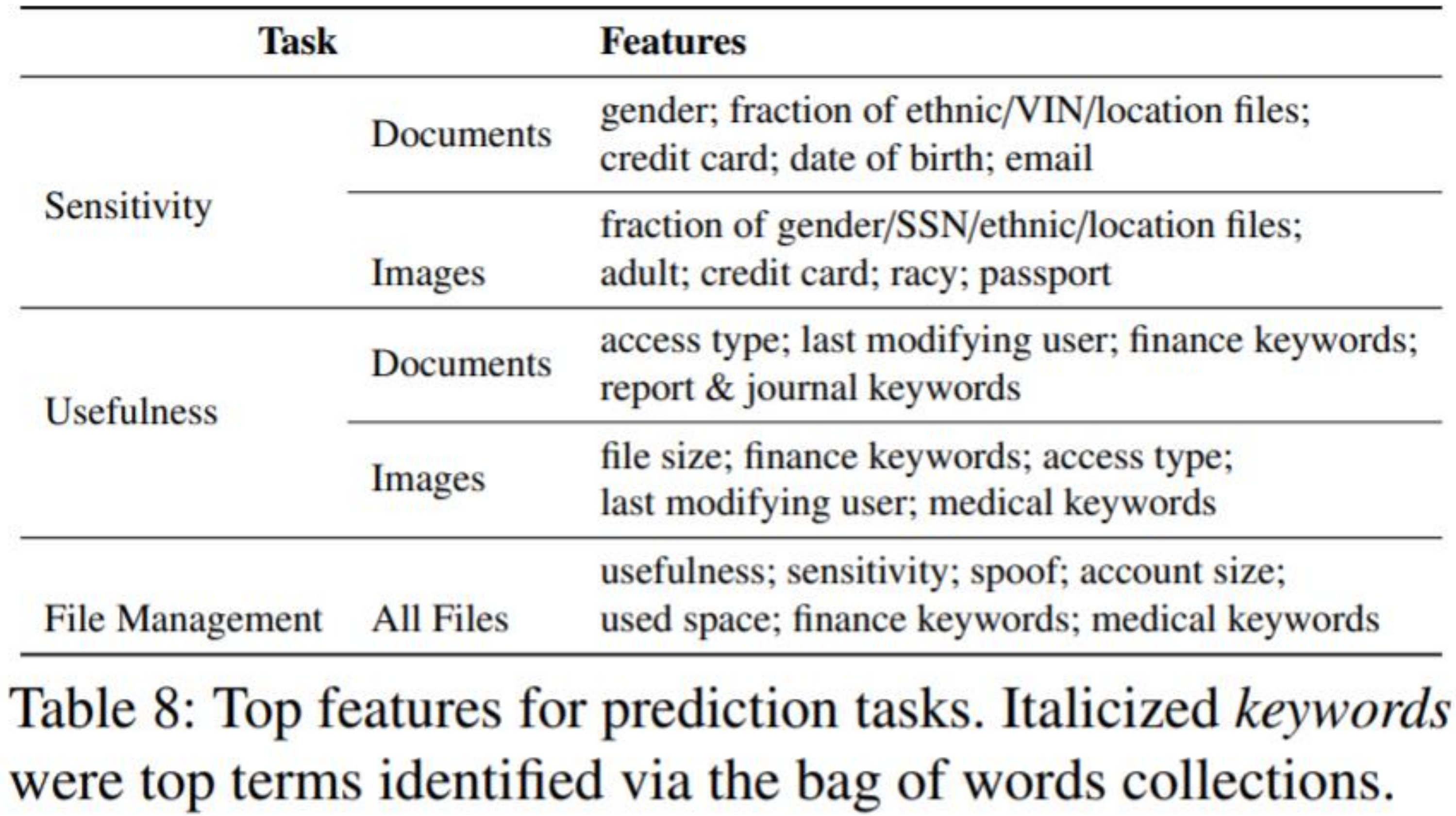


Category	Collection Method	List of Features
Metadata	Google Drive/Dropbox API	account size, used space, file size, file type (img, doc, etc.), extension (jpg, txt, etc.), last modified date, last modifying user, access type (owner, editor, etc.), sensitive filename, sharing status
Documents	Local text processing	bag of words for top 100 content keywords, LDA topic models, TF-IDF vectors, word2vec representations, table schemas for spreadsheets
Images	Google Vision API [20]	image object labels, adult, racy, medical, violent, logos, dominant RGB values, average RGB value
Sensitive Identifiers	Google DLP API [18]	***counts*** of the following identifiers in a file: name, gender, ethnic group, address, email, date of birth, drivers license #, passport #, credit card, SSN, bank account #, VIN

Table 3: A list of the features we automatically collected for each file using multiple APIs and custom code.

Taken from Khan et al. "Helping Users Automatically Find and Manage Sensitive, Expendable Files in Cloud Storage," USENIX Security 2021. https://www.blaseur.com/papers/usenix21-aletheia.pdf

Example of (Globally) Important Features



Task		Features
Sensitivity	Documents	gender; fraction of ethnic/VIN/location files; credit card; date of birth; email
	Images	fraction of gender/SSN/ethnic/location files; adult; credit card; racy; passport
Usefulness	Documents	access type; last modifying user; finance keywords; report & journal keywords
	Images	file size; finance keywords; access type; last modifying user; medical keywords
File Management	All Files	usefulness; sensitivity; spoof; account size; used space; finance keywords; medical keywords

Table 8: Top features for prediction tasks. Italicized *keywords* were top terms identified via the bag of words collections.

Taken from Khan et al. "Helping Users Automatically Find and Manage Sensitive, Expendable Files in Cloud Storage," USENIX Security 2021. https://www.blaseur.com/papers/usenix21-aletheia.pdf

Pitfalls of Feature Engineering

- ML model performance depends on the input data
  - Is the training data representative of the population?
  - Are the transformations applied to the data correct?
  - Is there enough training data to learn a good model?
- Many potential pitfalls throughout the process
  - Even careful humans will make mistakes!
- AutoML and automatic augmentation techniques
  - Opportunity or threat?

Model Selection

- Backward elimination
    - Start with all variables and eliminate one by one
- Forward selection
    - Start with no variables and add one by one

Training Data

Training Datasets and Benchmarks

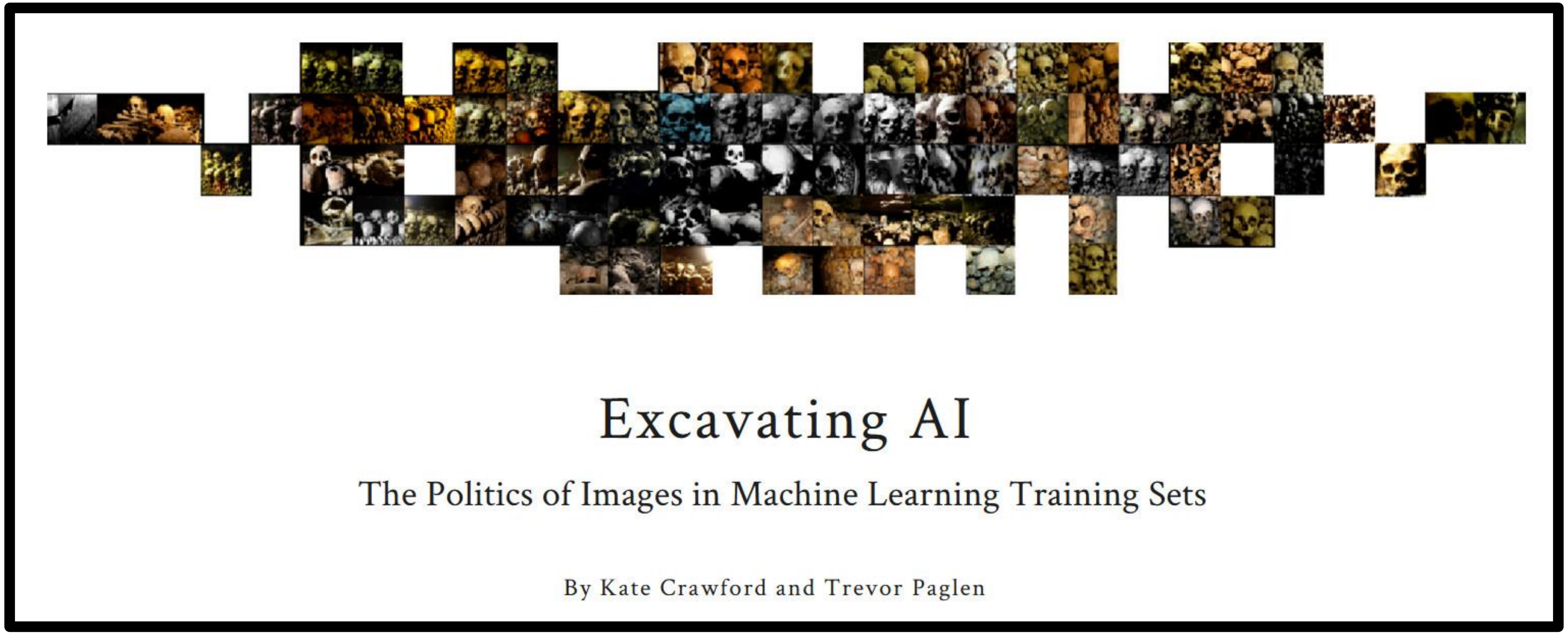
- Standardization of training datasets and benchmarks have arguably pushed the field of ML forward
    - Not without pitfalls
- If everyone is testing against the same datasets, what does that say about the ML model's generalizability?
    - Are results practically significant?
    - Do we notice errors that occur for data **excluded** from reference sets?
- There are more serious problems than a lack of progress!

Imagenet: Computer Vision dataset

- 15 million images
    - Each image is annotated with a noun from Wordnet
        - Wordnet -> hierarchy of concepts
- Instrumental dataset to advance computer vision
- Where did these images come from?

What Datasets Include/Exclude

- *Kate Crawford and Trevor Paglen, “Excavating AI: The Politics of Training Sets for Machine Learning (September 19, 2019)*
- https://excavating.ai



Excavating AI

The Politics of Images in Machine Learning Training Sets

By Kate Crawford and Trevor Paglen

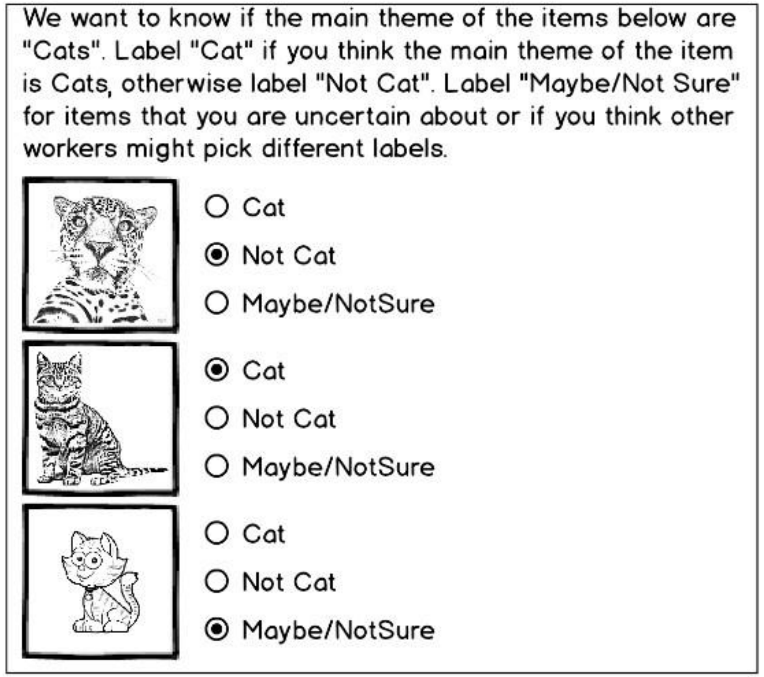
What Datasets Include/Exclude

- “the automated interpretation of images is an inherently social and political project, rather than a purely technical one”
- “What work do images do in AI systems? What are computers meant to recognize in an image and what is misrecognized or even completely invisible?”
- “how do humans tell computers which words will relate to a given image? And what is at stake in the way AI systems use these labels to classify humans, including by race, gender, emotions, ability, sexuality, and personality?”
- “As the fields of information science and science and technology studies have long shown, all taxonomies or classificatory systems are political.”

What Datasets Include/Exclude

“There is much at stake in the architecture and contents of the training sets used in AI. They can promote or discriminate, approve or reject, render visible or invisible, judge or enforce. And so we need to examine them—because they are already used to examine us—and to have a wider public discussion about their consequences, rather than keeping it within academic corridors. As training sets are increasingly part of our urban, legal, logistical, and commercial infrastructures, they have an important but underexamined role: the power to shape the world in their own images.”

Where Did the Labels Come From?



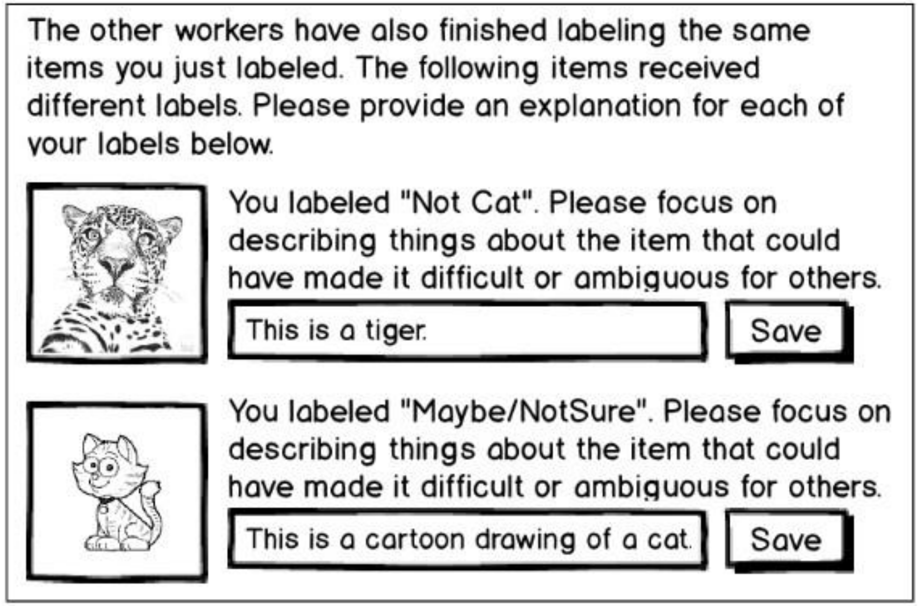
We want to know if the main theme of the items below are "Cats". Label "Cat" if you think the main theme of the item is Cats, otherwise label "Not Cat". Label "Maybe/Not Sure" for items that you are uncertain about or if you think other workers might pick different labels.

○ Cat
● Not Cat
○ Maybe/NotSure

● Cat
○ Not Cat
○ Maybe/NotSure

○ Cat
○ Not Cat
● Maybe/NotSure

Figure 3. Human Intelligence Task (HIT) interface for the Vote Stage. In addition to the predefined labels, crowdworkers can also select *Maybe/NotSure* when they were uncertain about the item.



The other workers have also finished labeling the same items you just labeled. The following items received different labels. Please provide an explanation for each of your labels below.

You labeled "Not Cat". Please focus on describing things about the item that could have made it difficult or ambiguous for others.

This is a tiger.

You labeled "Maybe/NotSure". Please focus on describing things about the item that could have made it difficult or ambiguous for others.

This is a cartoon drawing of a cat.

Figure 4. Human Intelligence Task (HIT) interface for the Explain Stage. Crowdworkers enter a short description for each item that was labeled differently in the Vote Stage. They were informed that disagreement occurred, but not the distribution of different labels used.

Chang et al. Revolt: Collaborative Crowdsourcing for Labeling Machine Learning Datasets. CHI 2017
https://dl.acm.org/doi/pdf/10.1145/3025453.3026044

Metrics

Some Possible Metrics (Classes)

- Accuracy: # correct / # total
- Confusion matrix (TP/FP/TN/FN)
- Area under the ROC curve (AUC)
  - True Positive Rate (TPR) = TP / P = TP / (TP + FN)
  - False Positive Rate (FPR) = FP / N = FP / (FP + TN)
  - ROC curve plots TPR vs. FPR at various thresholds
- Precision: TP / (TP + FP)
- Recall: TP / (TP + FN)
- Precision-Recall Curve

See https://medium.com/analytics-vidhya/performance-metrics-for-machine-learning-models-80d7666b432e
https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_algorithms_performance_metrics.htm
https://www.justintodata.com/machine-learning-model-evaluation-metrics/ or many more!

Some Possible Metrics (Numbers)

- Mean Squared Error
- Mean Absolute Error

See https://medium.com/analytics-vidhya/performance-metrics-for-machine-learning-models-80d7666b432e
https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_algorithms_performance_metrics.htm
https://www.justintodata.com/machine-learning-model-evaluation-metrics/ or many more!

Some Possible Metrics **Revisited**

- Do these metrics capture the **relationship** between **errors?**
- Do these metrics capture the **impact of errors?**
- Do these metrics capture the **differential** impact of **particular types of errors?**
- (Setting the stage for our next lecture!)

Some Possible Metrics (Performance)

- Model training time
- Frequency of model re-training
- Model size
- Classification time
- Privacy issues of the model
- “Security” (future lecture)

Commoditization of ML

ML Models as a Commodity

- We’ve talked about ML as:
  - Find a training dataset, goal, metric
  - Train the model
  - Use it for the task at hand
- Many models take many weeks to train in data-center scale computers. They are made available to everyone publicly:
  - Download off-the-shelf models
  - They have been trained with data that may not be available to you

The Trend Continues

- Models used as part of services packaged in the cloud vendors
    - Example: parts of AutoML offerings
- It’s easy to lose sight of the models you are using
- Many models are an amalgam of others: e.g., consider NLP
    - Input data is *featurized* using a ML model
        - GPT-3, BERT, Transformer-like models
    - Parameters may have been *pre-trained* with some dataset
    - Then you *fine-tune* to your data
- Allen NLP examples
- Kaggle

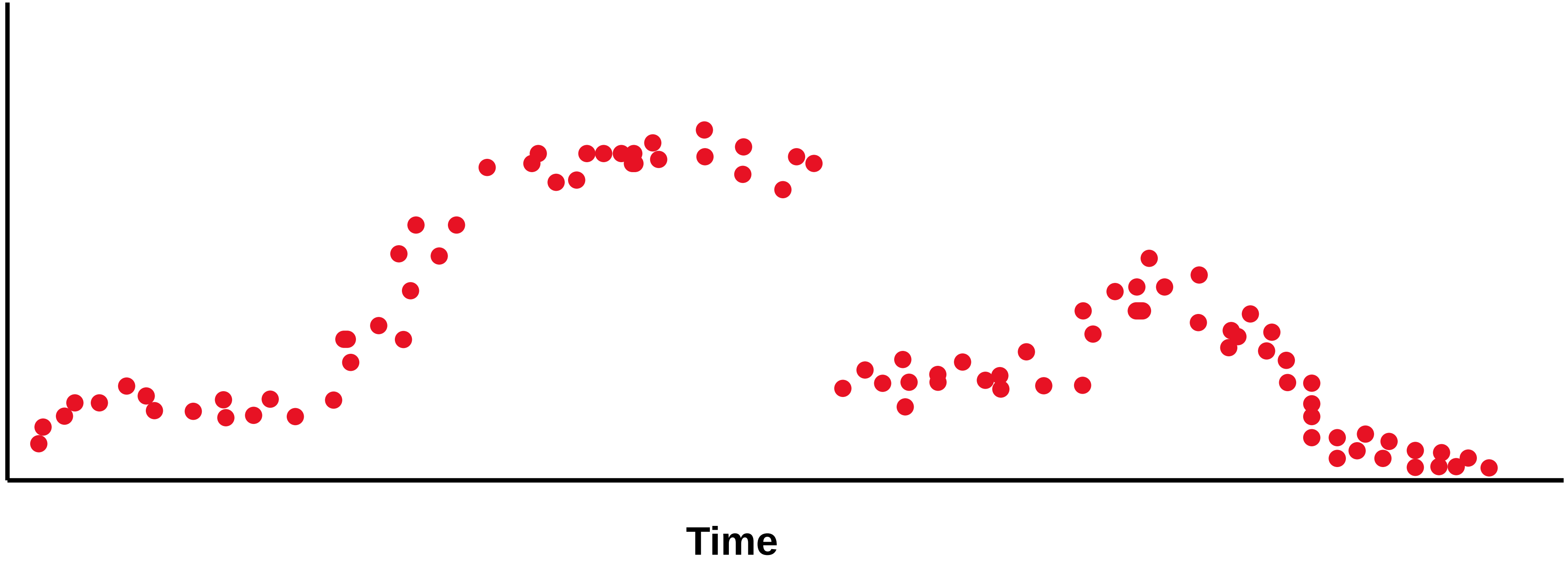
ML in the Wild

Concept Drift

ML in Pipelines

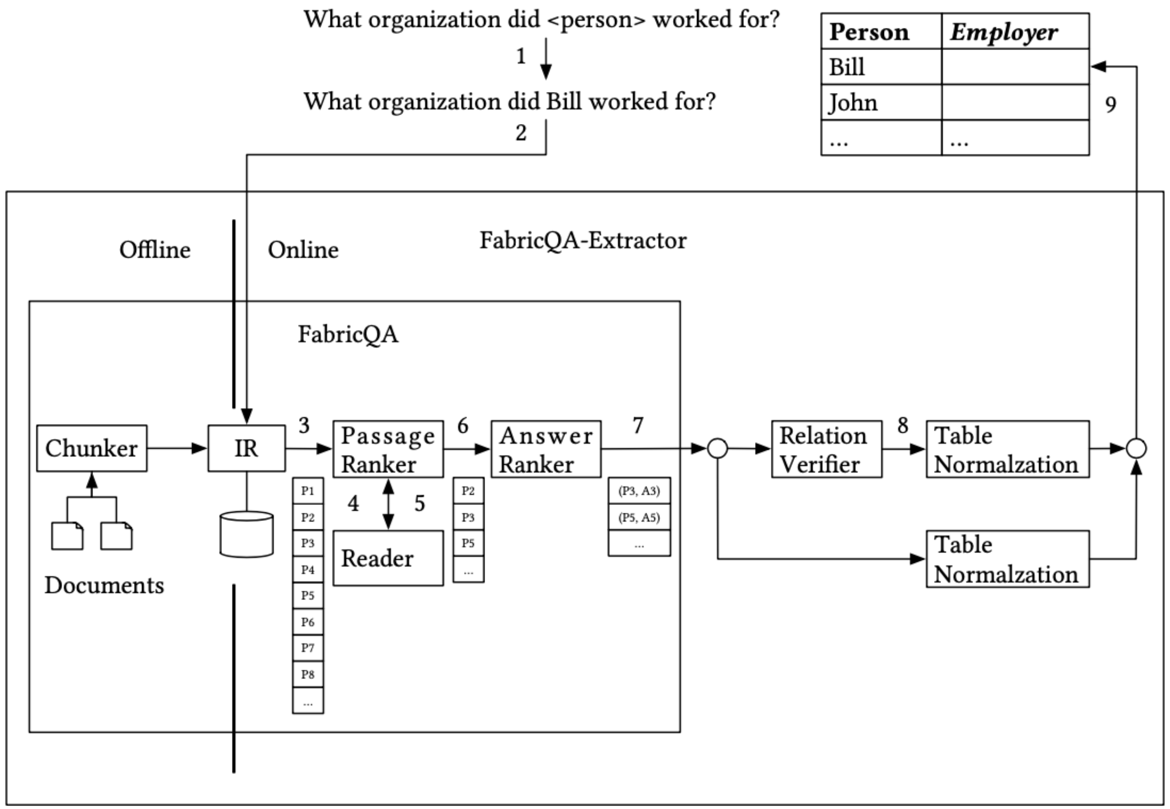
Concept Drift – The Passage of Time

- Extrapolation and Generalization
  - What population does the training data represent? At what point?
  - What claim can we make about the result?



Real systems use multiple models

Example: An information extraction system



Person	*Employer*
Bill	
John	
...	...

Figure 5: System Architecture

Algorithmic Decision Making

The Application Context Matters Greatly

Hiring

Online Advertising

Student Admissions

Criminal Justice

Health Insurance Markets

Creditworthiness

Selbst et al.’s Five Pitfalls

- Framing Trap
  - “Failure to model the entire system over which a social criterion, such as fairness, will be enforced“
- Portability Trap
  - “Failure to understand how repurposing algorithmic solu- tions designed for one social context may be misleading, inaccurate, or otherwise do harm when applied to a different context”
- Formalism Trap
  - “Failure to account for the full meaning of social concepts such as fairness, which can be procedural, contextual, and contestable, and cannot be resolved through mathematical formalisms”
- Ripple Effect Trap
  - “Failure to understand how the insertion of technology into an existing social system changes the behaviors and embedded values of the pre-existing system”
- Solutionism Trap
  - “Failure to recognize the possibility that the best solution to a problem may not involve technology”

Selbst et al. Fairness and Abstraction in Sociotechnical Systems. FAT*, 2019. https://dl.acm.org/doi/pdf/10.1145/3287560.3287598

What Does Accountability Mean Here?

- Who’s accountable for the consequences of an ML model?
  - Those who deployed it?
  - Those who built it and trained it?
  - The owners of the training data?
  - Those who listened to the algorithm?