Machine Learning in the Wild

Hiring

Online Advertising

Student Admissions

Criminal Justice

Health Insurance Markets

Creditworthiness

The Ingredients of an ML application

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

Outline

- Feature Engineering
  - Information leakage
- Training data
- Commoditization of ML
  - Reuse of datasets and models
- ML in the Wild
  - Concept Drift
  - ML in Pipelines
- Algorithmic Decision Making
  - ML as an artifact within Sociotechnical systems

Feature Engineering

Feature Representation

- Table data -> Matrix
- Row -> vector
- Transform categorical variables into a numerical representation
- Normalization, standardization, binning, and other transformations

Let's Build a Model To Understand Data

- Running example: a regression problem
- Example:

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 $\quad \hat{y}$

Variables/Attributes/Columns become 'features' of the input vector

Feature Engineering 1/2

- Feature Engineering / Model Selection
- Goal: To select the variables to feed into the model
  - More variables not always lead to better models

Feature Engineering 1/2

- Feature Engineering / Model Selection
- Goal: To select the variables to feed into the model
  - More variables not always lead to better models
- Backward elimination
  - Start with all variables and eliminate one by one
- Forward selection
  - Start with no variables and add one by one

Augmenting Features 2/2

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

Name	Age	Department
Jack	55	CS
Jane	27	Stats

JOIN

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

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

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?

What features are you selecting?

- How can you be sure that sensitive information is not represented in the model?
  - Is removing protected attributes enough?
  - Think about information leakage
- Anonymizing PII information
  - 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	??

Outline

- Feature Engineering
  - Information leakage
- Training data
- Commoditization of ML
  - Reuse of datasets and models
- ML in the Wild
  - Concept Drift
  - ML in Pipelines
- Algorithmic Decision Making
  - ML as an artifact within Sociotechnical systems

Training Data

Training datasets and Benchmarks

- Standardization of training datasets and benchmarks have arguably pushed the field of ML forward
  - No without pitfalls
- If everyone is testing against the same datasets, what does that say about the ML model's generalizability?
  - Are results practically significant?
- There are more serious problems than lack of progress

Imagenet: Computer Vision dataset

- 15M images
  - Each image is annotated with a noun from Wordnet
    - Wordnet -> hierarchy of concepts
- Instrumental dataset to advance computer vision

What’s in a training dataset?

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

What's in a training dataset?

- "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."

“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.”

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
  - E.g., 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, "*Markets for Models*", etc.

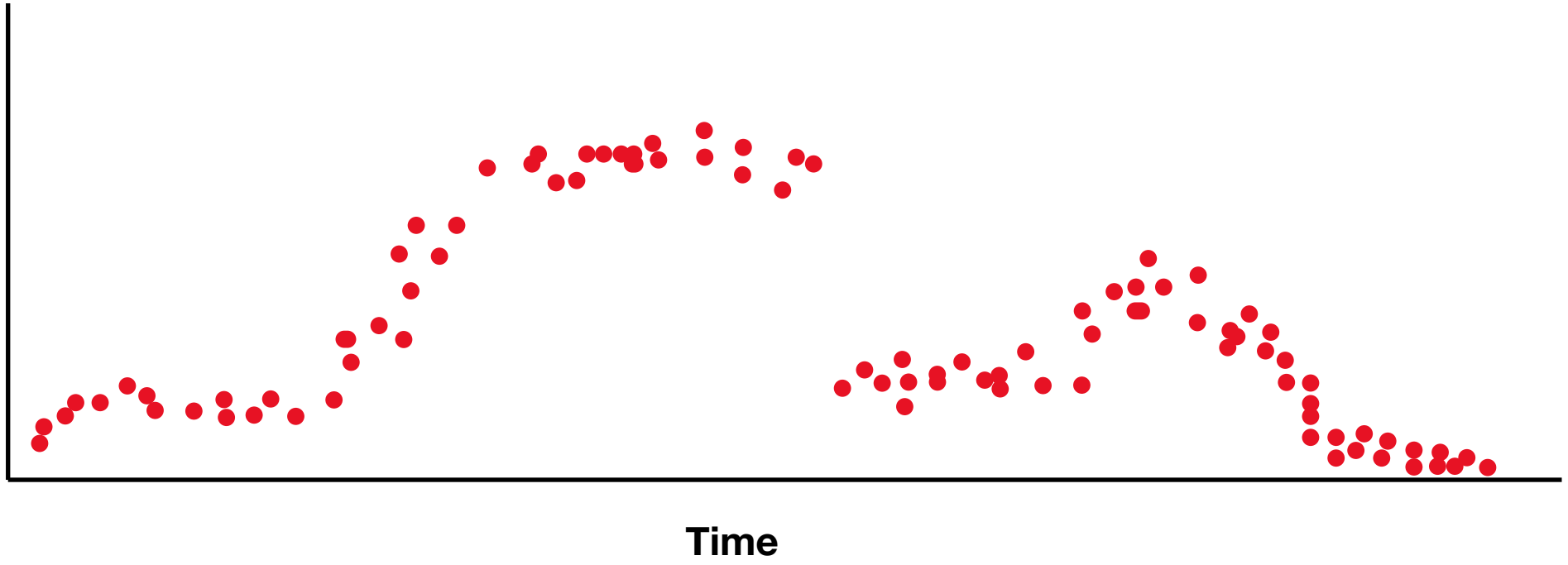
ML in the Wild

Concept Drift

ML in Pipelines

Concept Drift – The Pass of Time

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



Real systems use multiple models

Example: An information extraction system

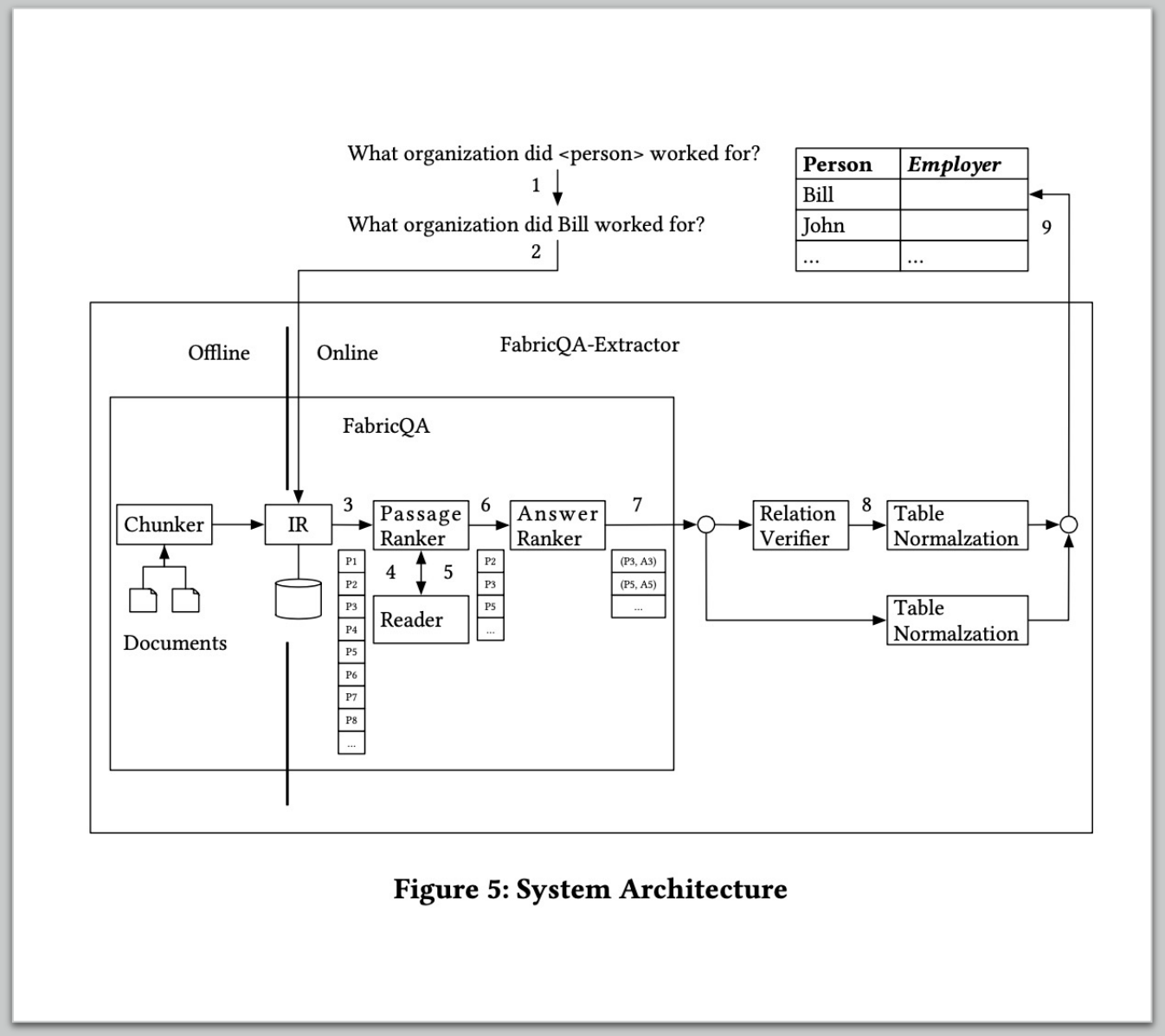


Figure 5: System Architecture

How do you debug those systems?

- “The AI revolution hasn’t happened yet” Michael Jordan

Algorithmic Decision Making

ML as part of Socio-Technical Systems

- Some external data fits a ML model
- The ML model produces an inference
- The inference is used to make/inform a decision

The 5 Pitfalls Selbst et. al. FACCT'19

- Framing Trap
  - "Failure to model the entire system over which a social criterion, such as fairness, will be enforced"
  - E.g., in ML, training data and labels
- 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"
  - E.g., the commodification of ML
- 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"
  - E.g., procedural vs outcome-based
- 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"

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?