Gov 2018 Part II: Machine Learning with Unstructured Data

Connor T. Jerzak

Visiting Asst. Professor ∗ Dept. of Government ∗ Harvard

Asst. Professor ∗ Dept. of Government ∗ UT Austin

Why ML with Unstructured Data?

- **Support causal inference**
  - Causal heterogeneity with high-dimensional *covariates*
  - Effect of high-dimensional *treatments*
  - Raw data sources as *proxy* for causal nodes
    - E.g., learning outcome representations from *raw data*
- **Support prediction**
  - Prediction itself sometimes important
    - Gov't instability, bail decisions, policy targeting
  - Prediction can support descriptive research
    - Summarizing massive data corpora
- **Improve welfare**
  - Policy action $\rightsquigarrow$ learn optimal actions in complex data envs.
- **Describe social science data better**
  - Social world $\rightsquigarrow$ incredibly complex
  - Our data $=$ static, researcher-created (e.g., surveys)
  - ML $\rightsquigarrow$ Gen. useful representations of complex data

ML with Unstructured Data

- **Some data have indefinite dimensionality**
  - Not only: *"More variables than data points"*
  - But:*"# of variables highly dependent on data rep."*
- EXAMPLE: **Text as indefinite dimensional object**
  - Document as bag of words:
    $w \in \mathbb{N}_0^{n_{\text{Words}}}$
  - Document as array of word embeddings:
    $w \in \mathbb{R}^{n_{\text{Words per Doc}} \times D_{\text{Embed}}}$
  - Document as array of character embeddings:
    $w \in \mathbb{R}^{n_{\text{Chars per Word}} \times n_{\text{Words per Doc}} \times D_{\text{Embed}}}$
  - Other examples: Image, video, audio, network, time series, etc.
- **Large-scale neural models & indefinite data**
  - *Neural nets:* Universal approx. theorems for continuous fxns
  - *Transformers:* Approximate Turning Complete systems
    * *With unstructured data:* We need higher levels of generic compute required to learn data representations along w/ outcome associations

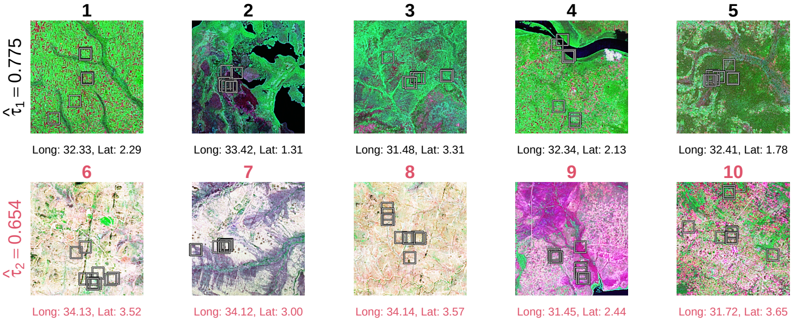
Integration of Software & Hardware

- *With unstructured data:* **We need higher levels of generic compute required to learn data representations along w/ outcome associations**
  - Model Size Scaling: More parameters $\rightsquigarrow$ Better performance?
  - Data Scaling: More data $\rightsquigarrow$ Better performance?
  - Compute Scaling: More training time $\rightsquigarrow$ Better performance?
- **Computational considerations $\rightsquigarrow$ Thus essential to achieve state-of-art results**
  - Leveraging (Multi-)GPUs/TPUs
  - Mixed precision training
  - Accurate quantized training
  - Model fine-tuning
- **Part II of course will touch on some of these concepts with social science data**

Example Application

Image-based Treatment Effect Heterogeneity

- **Question**: How can we use medical/satellite images to learn about the kinds of people who respond differently to an intervention?
- **Data pipeline**:
  - Processed satellite image data from Landsat
  - Individual-level data from an experiment in Uganda
  - Approximate individual geo-locations
- **Modeling pipeline**:
  - Approximate Bayesian inference
  - Bayesian Convolutional Neural Network/Vision Transformer
  - Clustering model for treatment effect distributions
  - Discussion of regularization, interpretability
- **Challenges**: Data leakage (test information in training set?), missingness in geo-location matches, comparing results via image, via tabular covariates



Top. High probability cluster 1 images.
Bottom. High probability cluster 2 images.