CORNELL CENTER for
SOCIAL SCIENCES

Accelerates, enhances, and amplifies social science research at Cornell.

Introduction to Machine Learning

Remy Stewart & Angel Hwang
CCSS-ResearchSupport@cornell.edu

CCSS Research Support Code of Conduct

The Cornell Center for Social Sciences provides a welcoming environment for everyone embracing all backgrounds or identities. All instructors and attendees agree to abide by our community norms. We encourage the following behaviors in our workshops:

- Respect differing viewpoints and ideas
- Share your own perspectives and ask any questions
- Accept constructive criticism
- Use welcoming and inclusive language
- Show courtesy and respect for all instructors and attendees

If you believe that an instructor or attendee has violated the code of conduct, please report the violation to CCSS-ResearchSupport@cornell.edu. We take all reported incidents seriously.

Land Acknowledgement

Cornell University is located on the traditional homelands of the Gayogo̱hó:nǫ' (the Cayuga Nation). The Gayogo̱hó:nǫ' are members of the Haudenosaunee Confederacy, an alliance of six sovereign Nations with a historic and contemporary presence on this land. The Confederacy precedes the establishment of Cornell University, New York state, and the United States of America. We acknowledge the painful history of Gayogo̱hó:nǫ' dispossession, and honor the ongoing connection of Gayogo̱hó:nǫ' people, past and present, to these lands and waters.

Here are additional links for more on the history of Cornell's violent, colonial formation, the movement to return native lands, and about the AIISP program at Cornell.

Consider donating to the Gayogo̱hó:nǫ' sovereignty initiative here.

Outline

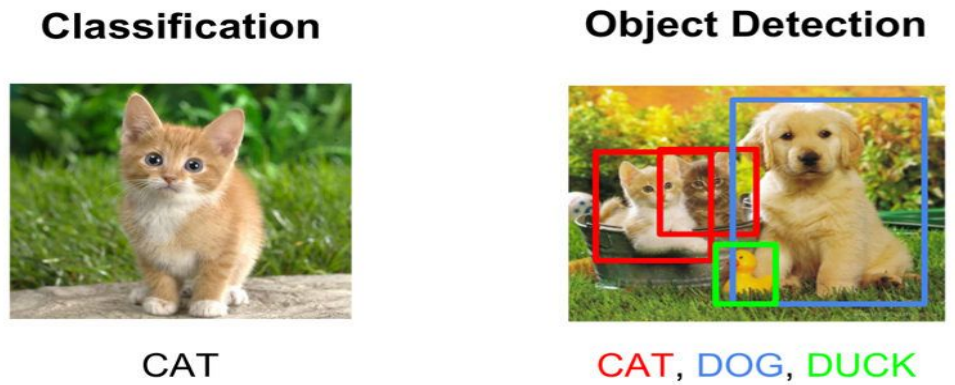
- **Introduction to Machine Learning**
- Understanding & Visualizing Data
- Supervised Learning
- Unsupervised Learning
- Natural Language Processing
- Conclusion

What is Machine Learning?

Class of statistical techniques that improve their performance on a task with experience, leading to data-informed predictions

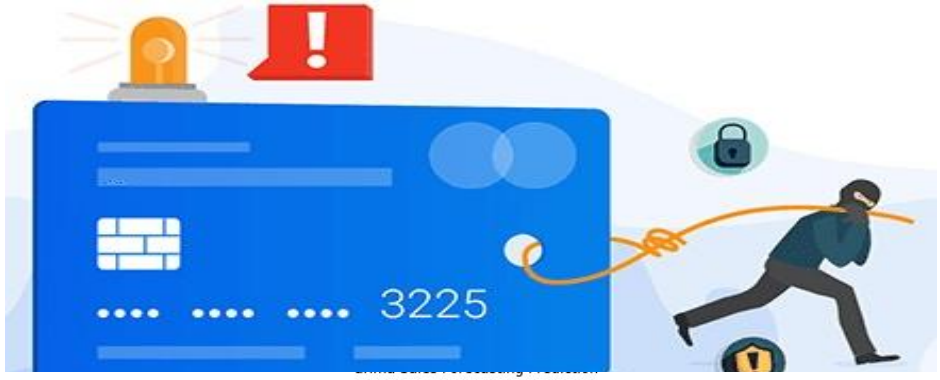
Identify patterns & generate replications

- By learning from thousands of images, a model can generate an accurate image of a “dog” or “cat”.



Flag abnormalities

- If fed a large collection of card transactions, a model can identify the rare occurrence of credit fraud.



Predict future outcomes

- Given the purchasing trends of customers over multiple years, a model can predict how much revenue to expect next month.

Unstructured vs. Structured Data

Structured Data

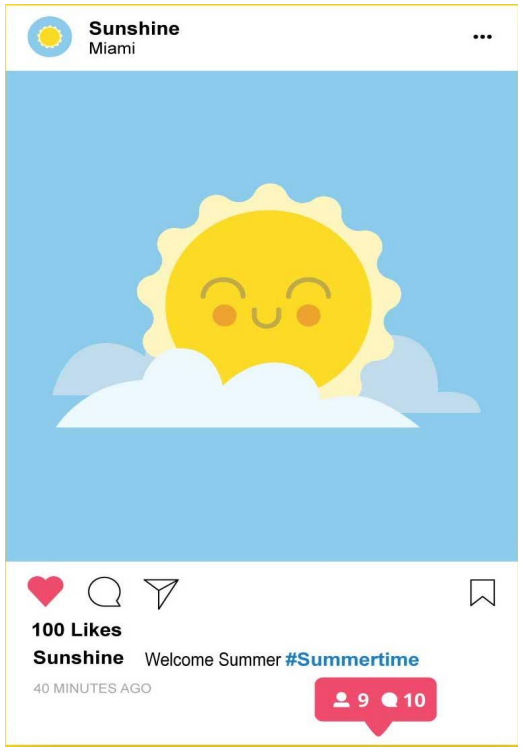
Pre-specified format: survey results with set columns for variables

B11

	A	B	C
1	**User**	**Satisfaction Score**	**Used integration**
2	Barbra C.	7	yes
3	Tom J.	9	no
4	Dana R.	8	no
5	Stacy M.	8	yes
6	Phil C.	9	yes
7	Steve M.	6	yes
8	Ciara G.	10	no
9			

Unstructured Data

No pre-specified format: raw text, images, audio, video



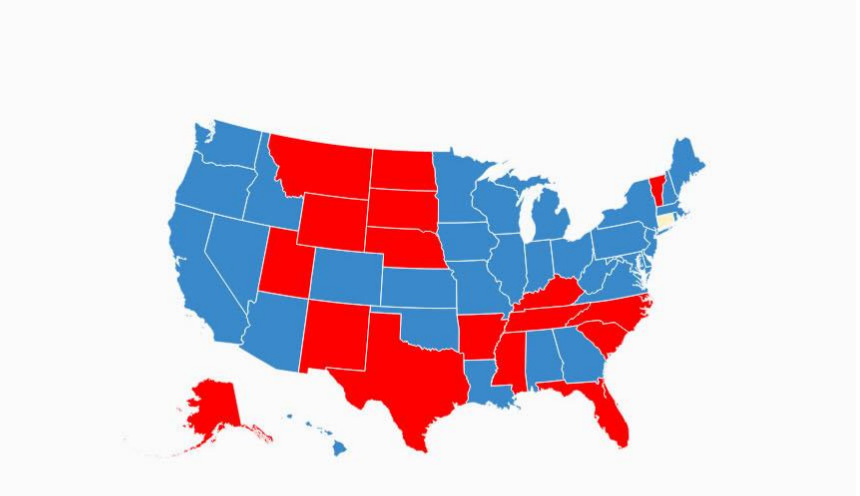
Why would social scientists be interested in ML?

- Apply innovative techniques to understand human behavior & social systems
- Can be integrated with classic subdiscipline theories & methodologies
- Applies to a variety of traditional social science data- panel surveys, digitized documents, experiments, etc.
- Well-suited for Big Data age and greater presence of social behavior now on digital platforms

Core tasks of discovery, measurement, causal inference, and prediction.

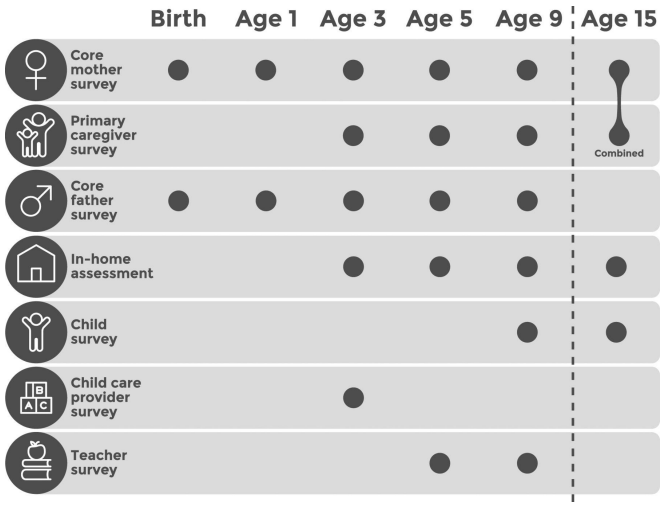
What questions can machine learning answer?

Supervised Learning:
Learning from labeled data



- I have collected millions of political tweets. How can I use these tweets to predict voting outcomes at the state level?

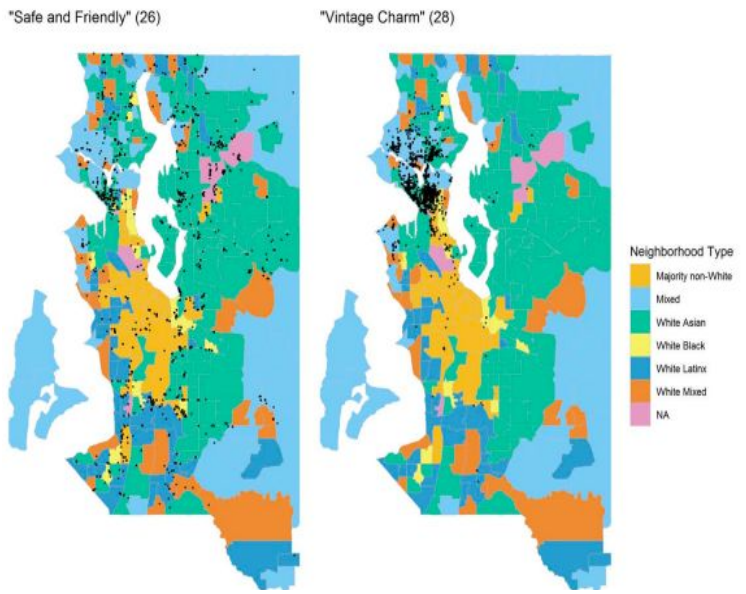
- I have a rich dataset regarding children and their families covering factors such as GPA and material hardships. How can I use this data to attempt to predict their life trajectories?

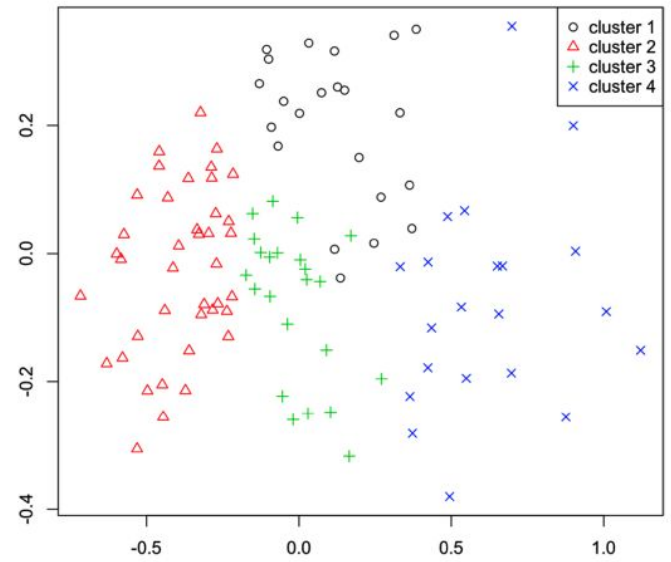


What questions can machine learning answer for?

Unsupervised Learning:
Learning from unlabeled data

- I have collected the text of rental advertisements from Craigslist. Does language to describe these rentals differ based on the neighborhood demographics of where the unit is located?
- I have survey data from managers of online businesses. How can I segment this market and learn the unique profiles of each segment?





Outline

- Introduction to Machine Learning
- **Understanding & Visualizing Data**
- Supervised Learning
- Unsupervised Learning
- Natural Language Processing
- Conclusion

Understanding Data through EDA

Exploratory Data Analysis (EDA) as the preliminary workflow to support informed decisions around handling data.

Data Pre-Processing for ML:

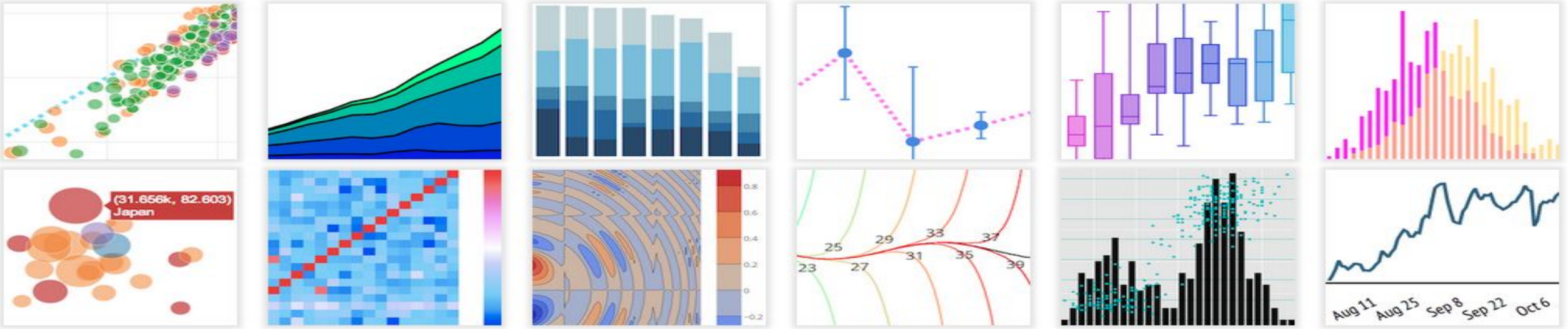
- Selecting, excluding, creating, and transforming variables
- Understanding relationships between independent and dependent variables
- Handling challenges around outliers, missing data, noisy data, etc.

Different Approaches Regarding:

- Type of ML model
- Structured vs. unstructured data
- Underlying research questions

Data Visualization

- Core component of EDA
- Guides interpretation beyond descriptive statistics
- Essential for interpreting & sharing ML findings
    - Classic graphs (histograms, scatter plots, box plots)
    - Novel visuals (heat maps, violin plots, ridge plots)



Outline

- Introduction to Machine Learning
- Understanding & Visualizing Data
- **Supervised Learning**
- Unsupervised Learning
- Natural Language Processing
- Conclusion

What is Supervised Learning?

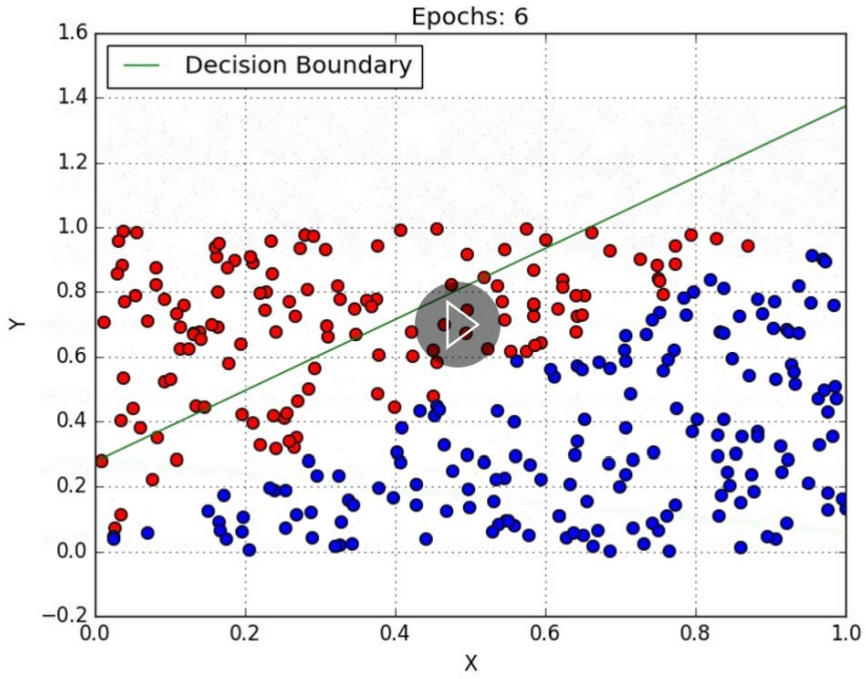
- Model the relationship between independent variables (model inputs) and dependent variables (model outputs) using a relatively small, labeled data.

- Use model for prediction with large dataset with unknown dependent variables.

- Research questions can be separated into *classification* tasks and *regression* tasks:
  - *Classification*: the labels are discrete
    - Binary: 0/1, Yes/No
    - Multi-class: Cat/Dog/Duck, Digit(0,1,..,9)
  - *Regression*: the labels are continuous
    - Wage, GPA, housing prices, etc.

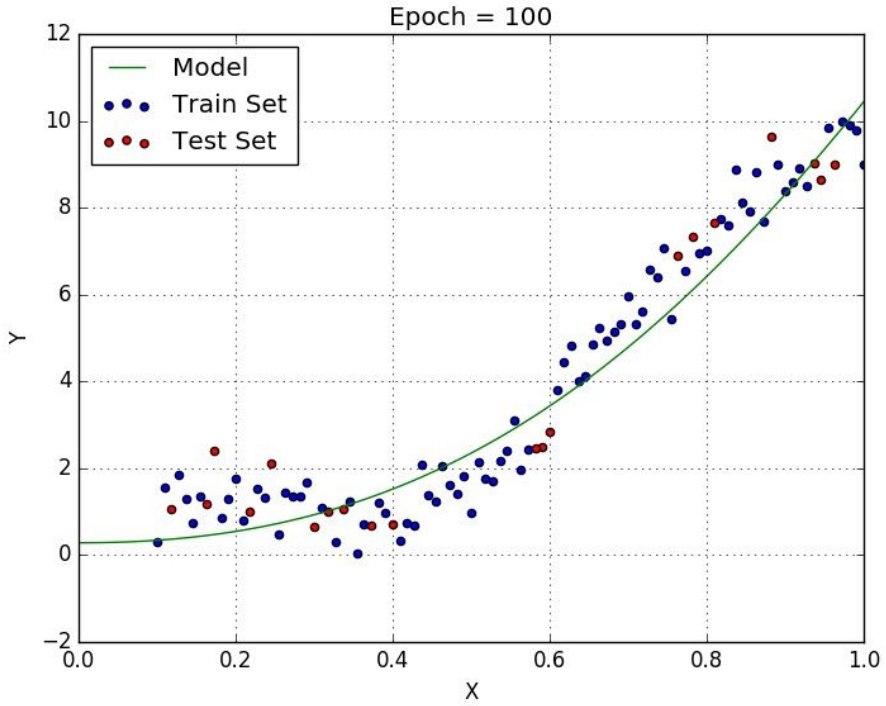
What is Supervised Learning?

Classification



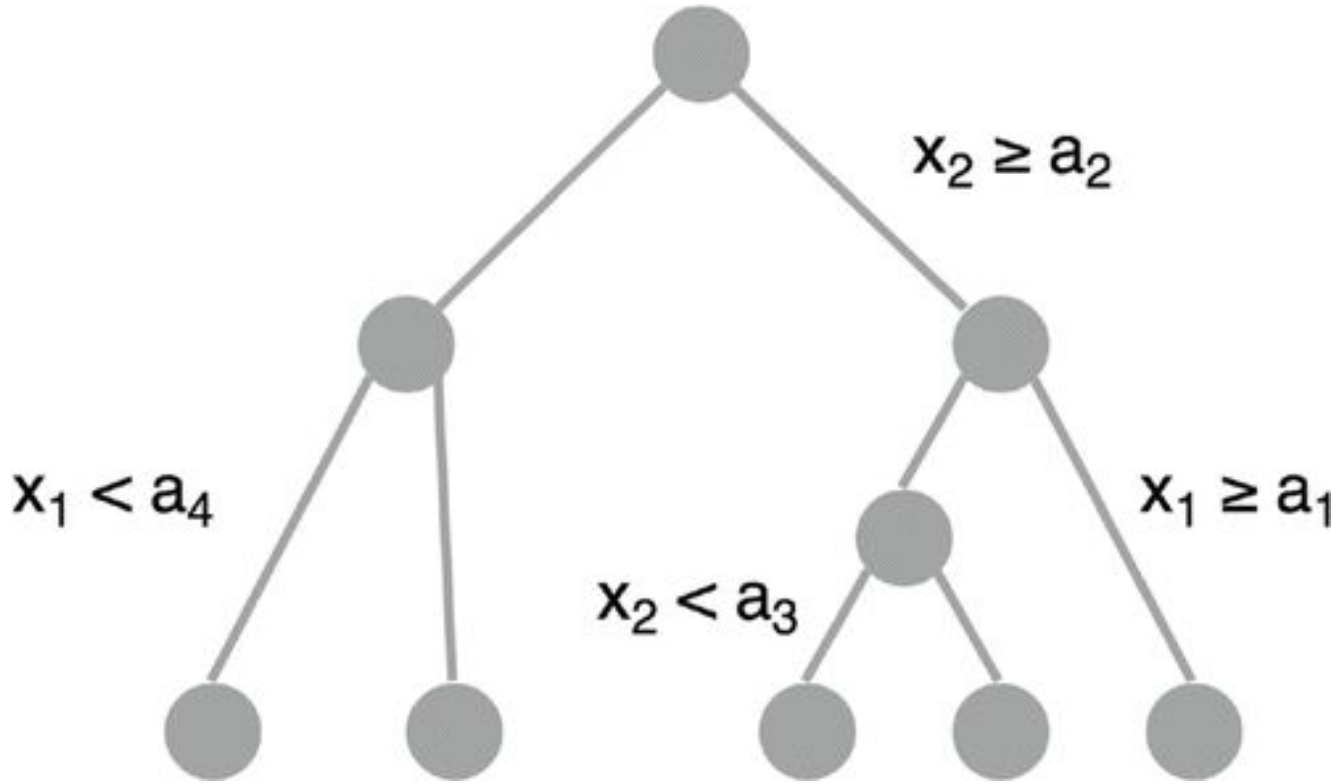
Source: Blog post by Davi Frossard

Regression



Source: Blog post by Davi Frossard

Tree-based Model



Source: Schonlau, M., & Zou, R. Y. (2020).

Supervised Learning Flowchart

Exploratory Data Analysis → Data Cleaning & Feature Engineering → Algorithm Selection → Model Training → Prediction & Interpretation

Understand the basic information of the dataset.

- Check distribution of the key variables.
- Check relationship between key variables.

Prepare the clean data for model fitting.

- Data Cleaning
  - missing values
  - outliers
  - data errors
- Feature Engineering
  - dummy variables
  - interaction terms
  - aggregating data
  - dropping redundant information

Select the models for model fitting.

- What is the practical benefits of each model?
- Should I adjust any model parameters?

Train the models to select the best model.

- Split data into training and test set.
- Fit the model using the training set.
- Select the best model based on model performance.

Make inference from the best model.

- Check the accuracy of the prediction.
- Explain the relationship between model inputs and outputs.
- Integrate models and explanations into theory building.

Outline

- Introduction to Machine Learning
- Understanding & Visualizing Data
- Supervised Learning
- **Unsupervised Learning**
- Natural Language Processing
- Conclusion

What is Unsupervised Learning?

- Learning from **unlabeled data**
- Given a set of features, finding patterns and structures in data
- Common techniques:
  - K-means (learning segments)
  - Hierarchical clustering, association rules (learning structural relationships)
- Unsupervised learning as a bottom-up approach to exploratory research

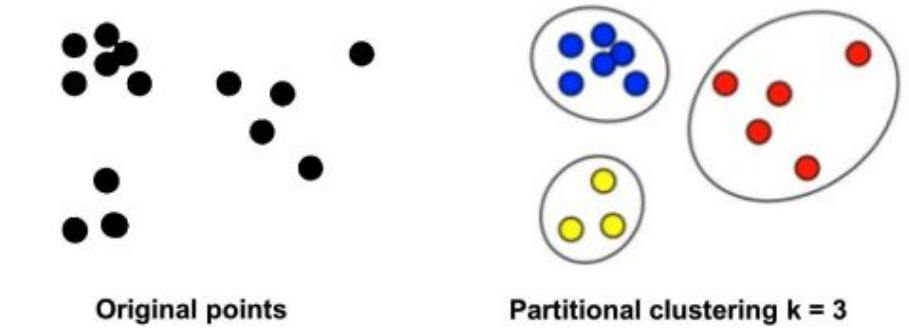


FIGURE 1. Example of clustering based on partitional methods.

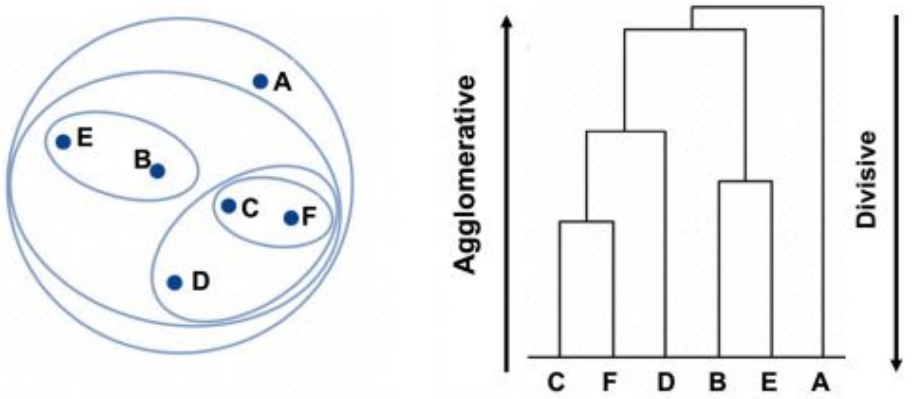


FIGURE 2. Example of clustering based on hierarchical methods.

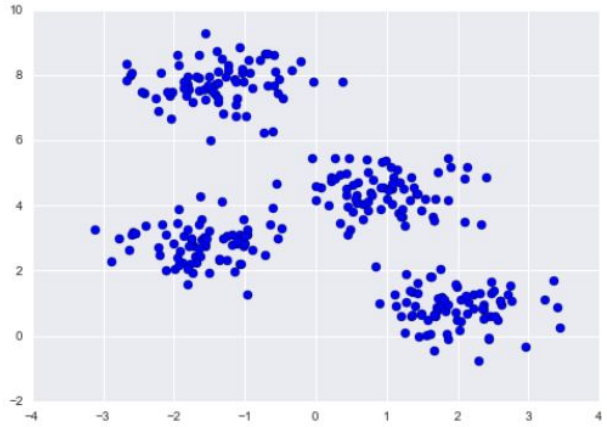
Luna-Romera et al., 2019

What is Unsupervised Learning?

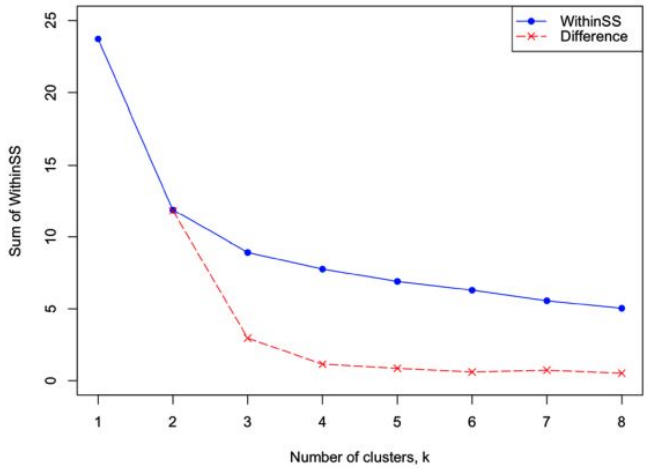
Clustering

Applying Unsupervised Learning as an Exploratory Tool

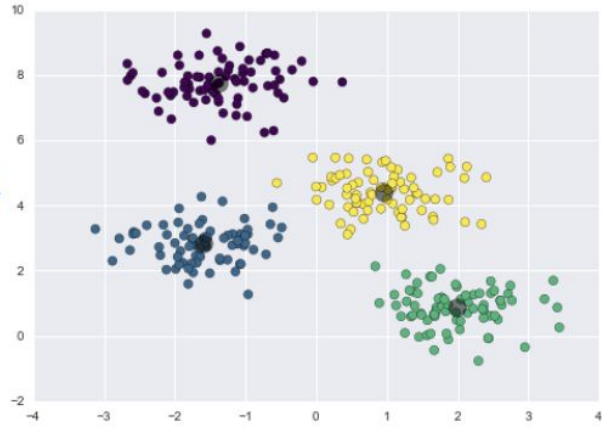
Explore & Visualize Data



Fit unsupervised learning models & determine optimal model parameters



Re-examine the quality of each newly learned “label”

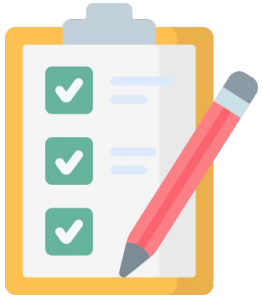


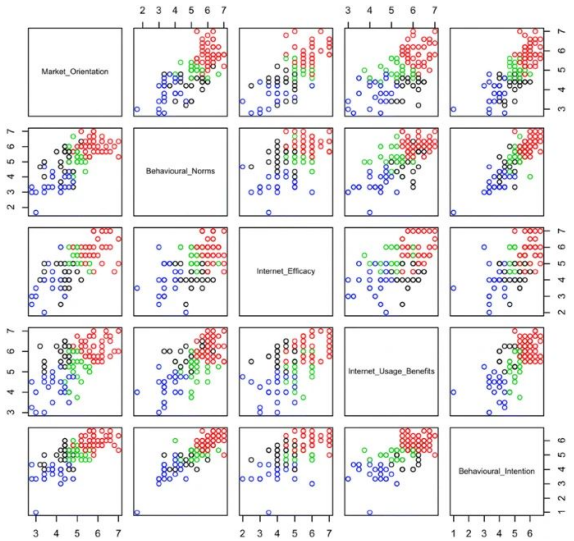
Applying Unsupervised Learning with Theoretical Support

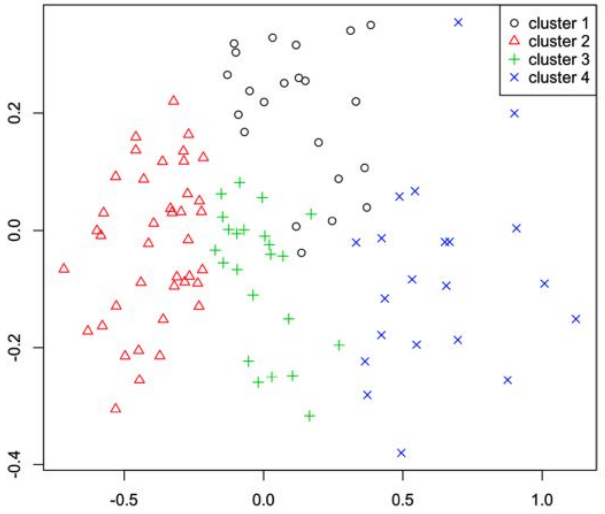
Identify important features based on existing work

Fit unsupervised learning models & determine optimal model parameters

Re-examine the quality of each newly learned “label”







Outline

- Introduction to Machine Learning
- Understanding & Visualizing Data
- Supervised Learning
- Unsupervised Learning
- **Natural Language Processing**
- Conclusion

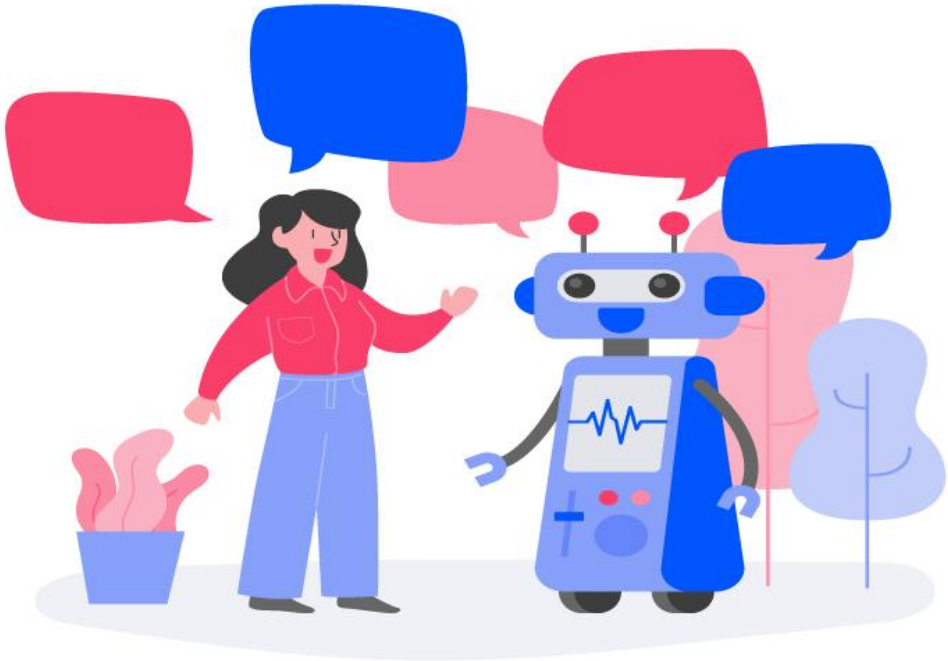
Natural Language Processing

Interdisciplinary field using computational tools to interpret text & language

Rapidly increasing amount of available text from online platforms & digitization of print

Intersection of data science tools with qualitative insights- “Computational Grounded Theory” (Nelson 2017)

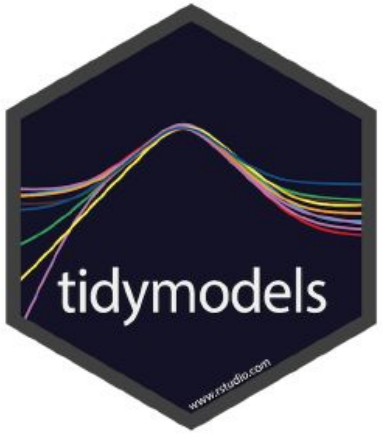
Well-suited application of ML for social scientists given our expertise in analyzing domain-specific text



Implementing NLP Research

Transformers

GENSIM
topic modelling for humans



spaCy

While not all NLP uses ML, some of the most innovative applications are ML-powered

Real-life applications within language translation, speech recognition (Siri/Alexa), transcription generation, text autocompletion, and beyond.

Growingly more accessible to implement across academic disciplines within both Python & R

Corpus collection with a total **vocabulary** consisting of a number of unique **documents** that each use a measurable combination of **word tokens**.

NLP Questions & Methods

Classification: How can group membership be identified by word use? How can we make predictions on unseen data by learning these language trends?

Clustering/Topics: What are the primary discussion points within a text corpus? Which subgroups are associated with certain topics?

Embeddings: How do we represent the core characteristics of a vocabulary at a lower dimensionality? How do words measure similar semantic concepts to each other?

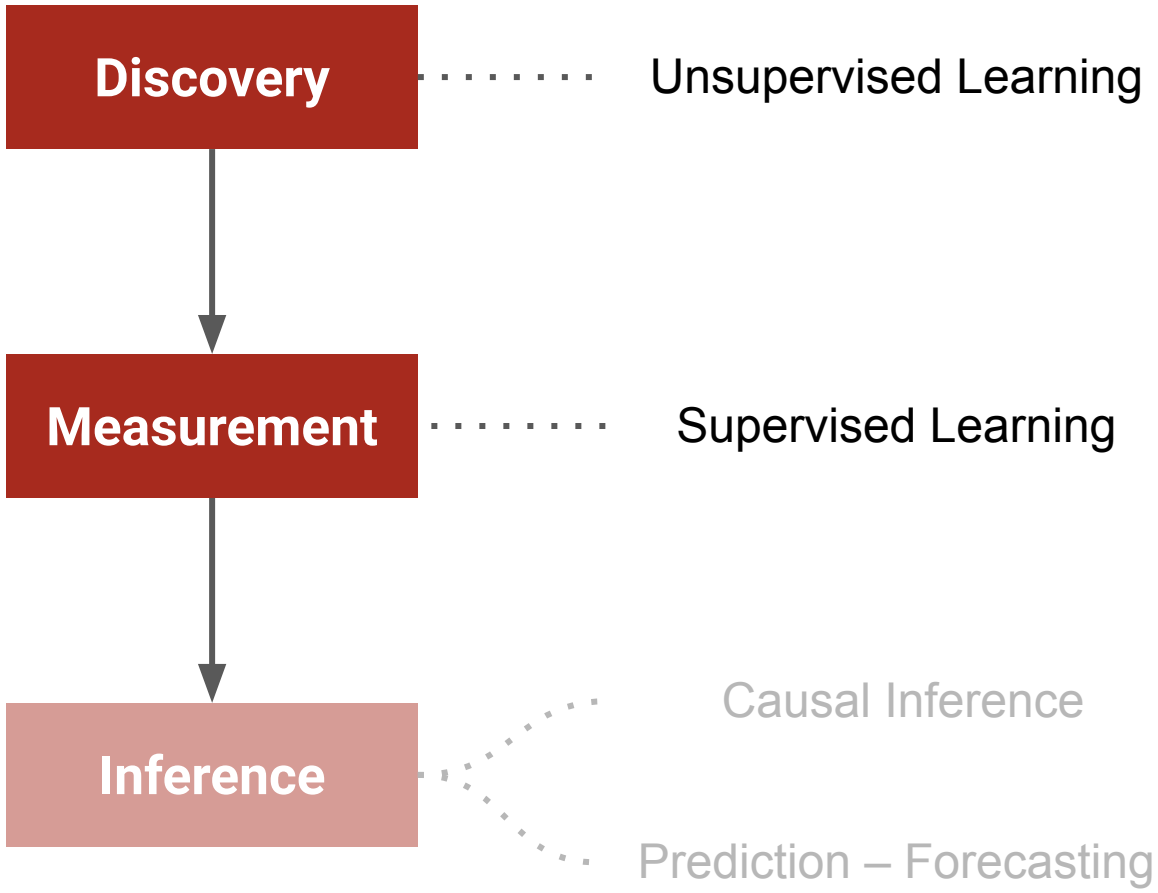
Networks: What are the word co-occurrence patterns across documents? How do text-based communities establish network ties?

Causality: How does changes in text lead to a change in outcomes? How can text be used to measure treatment effects?

Outline

- Introduction to Machine Learning
- Understanding & Visualizing Data
- Supervised Learning
- Unsupervised Learning
- Natural Language Processing
- **Conclusion**

When can social scientists apply machine learning in research?

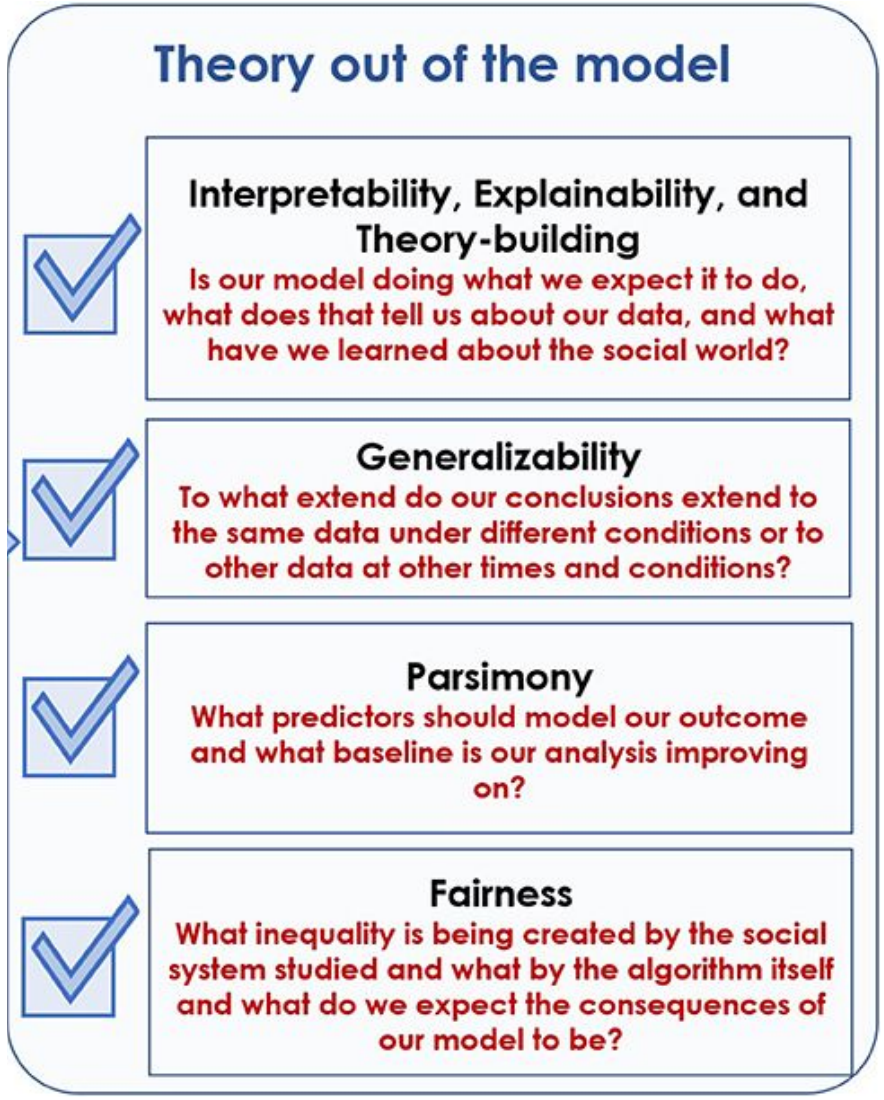


Other emerging potentials:

- Applying ML to facilitate and diminish biases in qualitative coding
- Social network analysis
- Visual & content analysis
- Analyzing audio/video content

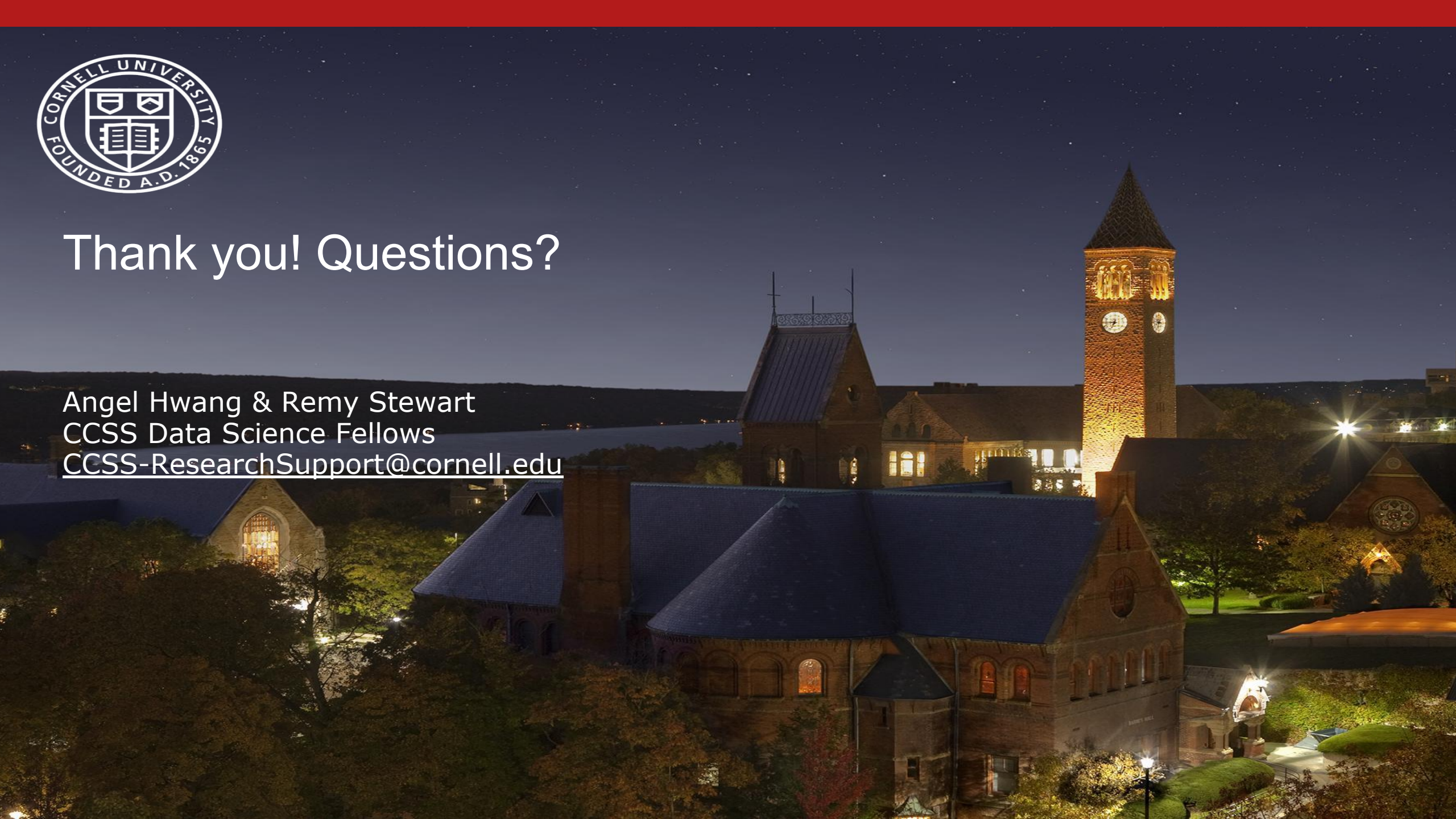
Adopted and revised from Grimmer et al., 2021

Wait… But is there any concern about ML in social science?



Source: Radford, J., & Joseph, K. (2020).

- Are the results interpretable and explainable? How accurately does the predicted output match what we expect? What is it about the (social) world that led to the model learning the relationships between inputs and outputs?
- Can we apply the same methodology to other data or different domains to see whether it performs similarly to the original?
- Are there social theories that support our selection on features / model inputs?
- Are models leading us to make discriminatory decisions and baseless social scientific claims?



Thank you! Questions?

Angel Hwang & Remy Stewart
CCSS Data Science Fellows
CCSS-ResearchSupport@cornell.edu

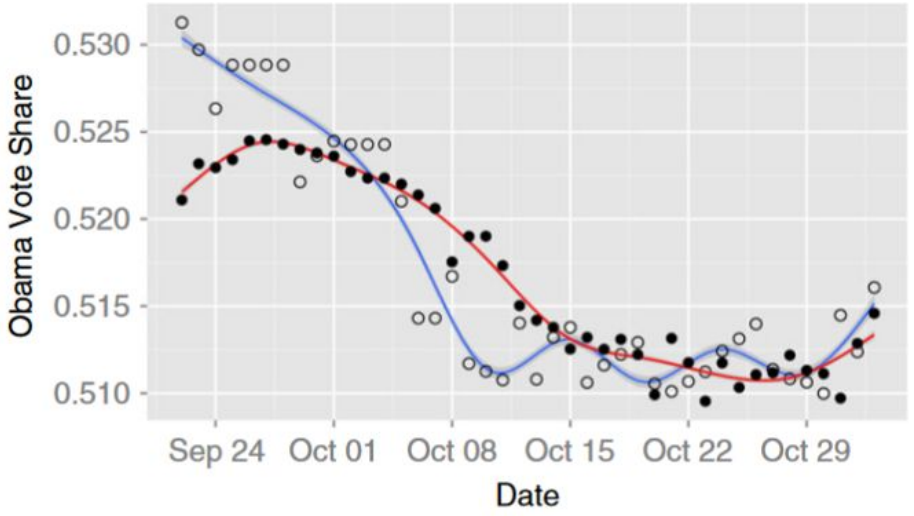
Supervised Learning Example 1

"Predicting and interpolating state-level polls using Twitter textual data"

Goal: With sufficiently large collection of political Twitter posts at state-day level, researchers can extrapolate vote intentions in state that are poorly polled, interpolate vote intentions for unpolled days, and measure quick changes in vote intentions even for well-polled states.

ML Tools: Predict average poll-measured vote intentions at state-day level using different **classification** models (regularization, random forest, support vector machine)

FIGURE 3 Predicted and Actual Polling for Ohio



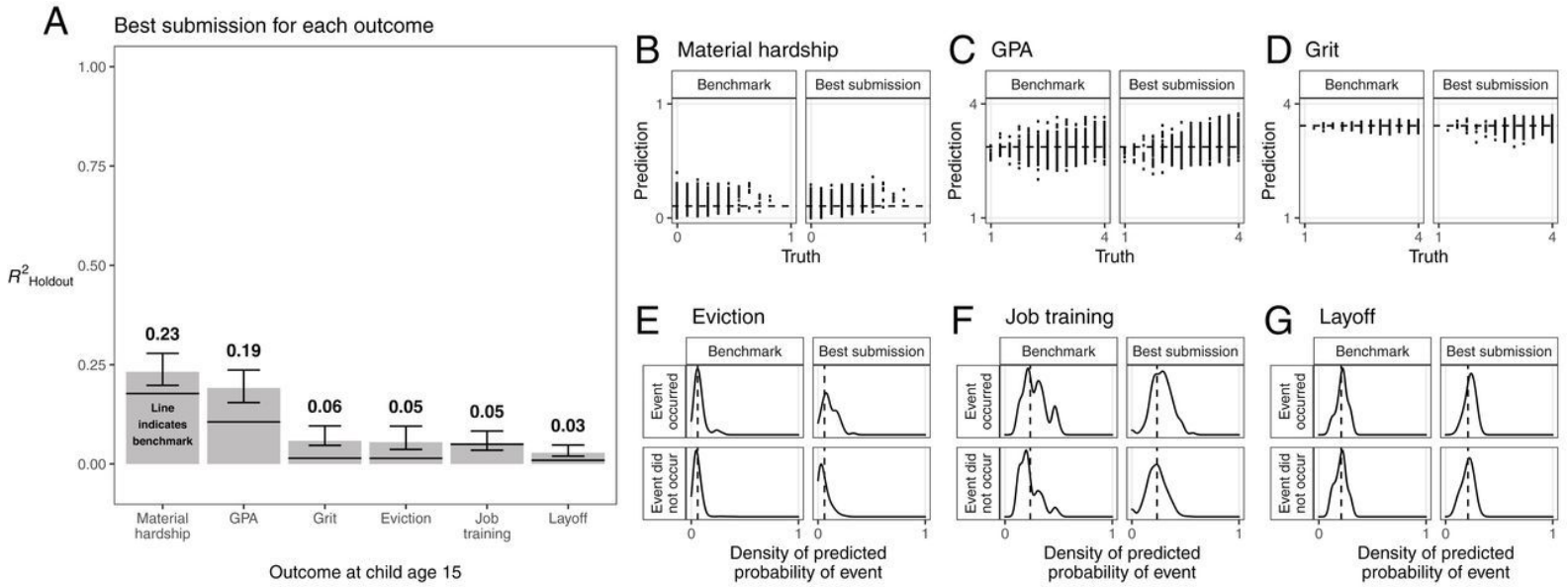
Note: Open circles indicate polls; filled circles indicate text-based predictions.

Beauchamp, N. (2017).

Supervised Learning Example 2

> "**Measuring the predictability of life outcomes with a scientific mass collaboration.**"

Goal: How predictable are life trajectories? 160 research teams investigated this question with a scientific mass collaboration using a rich dataset and building predictive models for six life outcomes (such as GPA and financial hardship). However, the best predictions were not very accurate and only slightly better than those from a simple benchmark model.



Salganik, M. J., Lundberg, I., Kindel, A. T., Ahearn, C. E., Al-Ghoneim, K., Almaatouq, A., ... & McLanahan, S. (2020).

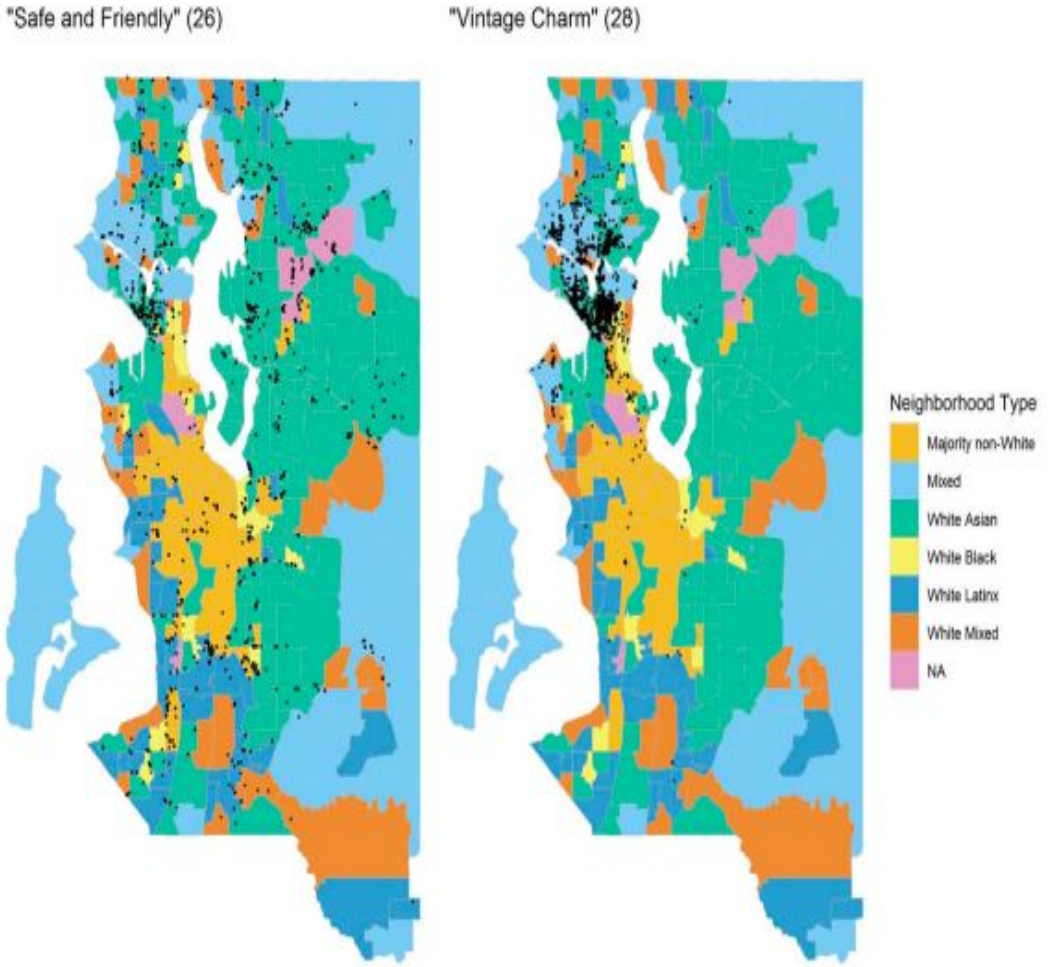
Unsupervised Learning Example 1

"Racialized Discourse in Seattle Rental Ads Text"

How does landlords' language change when advertising rentals in neighborhoods with different racial/ethnic demographics?

Fit structural topic models to measure language change by residential composition

- Majority White neighborhoods more likely to be described as desirable and historical communities
  - "vintage", "warm", "craftsmanship", "golf"
- Majority POC neighborhood descriptions focused on safety, development, and practical amenities
  - "convenient", "security", "income report", "emerging"



Kennedy, Hess, Paullada, & Chasins (2021)

Unsupervised Learning Example 2

Analyzing ecology of Internet marketing in small- and medium-sized enterprises (SMEs) with unsupervised-learning algorithm

How can we segment small-/mid-sized enterprises?
What does the profile of each segment look like?

- Informed by existing theories, the authors selected 5 dimensions to perform clustering, segmenting N = 106 enterprises into 4 clusters.
- The authors then quantitatively and qualitatively examined the characteristics of each segment.

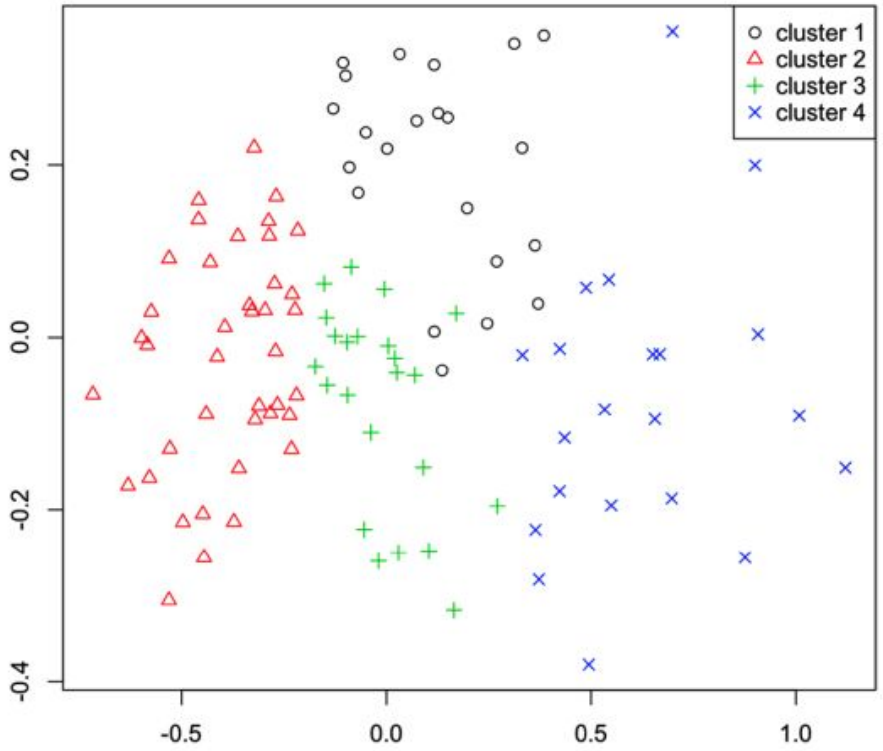


Table 4 The characteristics of four clusters

Clusters (symbol)	Sizes	Centers					Within SS
		Marketing orientation	Behavioral norms	Internet efficacy	Internet usage benefits	Behavioral intention	
1 (○)	22	0.35	0.64	0.45	0.72	0.70	1.61
2 (Δ)	40	0.73	0.83	0.78	0.79	0.88	2.49
3 (+)	23	0.53	0.68	0.67	0.56	0.73	1.31
4 (×)	21	0.25	0.34	0.38	0.30	0.48	2.36
Total	106	0.51	0.66	0.61	0.63	0.73	7.77

Yau & Tang (2018)