

# This Economist and Machine Learning

(De-mystifying the Un-mystical)

Tim Savage

# Machine Learning (ML) Is Statistical Learning

- Machine learning is simply stochastic modeling.
- Linear regression came first and remains an indispensable ML tool.
- Econometrics is a sizeable corner of the ML room.
  - Identification and treatment effects.
- But the array of problems amenable to data analysis has expanded greatly in the last decade or so.
  - For these problems, regression does not work well or at all.
  - The toolkit had to be expanded well beyond regression.
  - And it started with the use of the humble logit for spam.
- Economists have been at the table for decades.

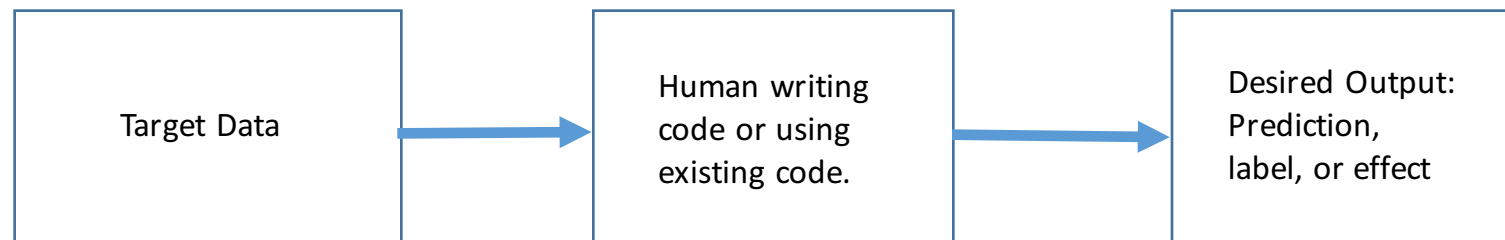
# If It Can Be Digitized, It Can Be Analyzed

- We are awash in the digital exhaust of human behavior.
- The contours of the explosion are obvious.
- Everything is saved because incremental cost is zero.
- Low barriers to entry due to open source engines like R and Python.
- Private sector sees potentially large returns to large-scale data mining (*e.g.*, recommendation systems for price discrimination).
- Public policy-makers say they want “data-driven decision-making”.
  - DARPA’s NGS2 (Next Generation Social Science) effort.
- ML tools work very well in practice and are improving quickly.

# The Irrelevant

- Computer scientists have been remarkably successful at creating a mystique around the phrase “machine learning”.
  - Bagging.
  - Multilayered perceptron and deep learning.
- And yet, “I am a data scientist” is now akin to “I play sports”.
- ML is not artificial intelligence, but many computer scientists consider it to be a branch of artificial intelligence.
- I believe computer scientists have won the “language war”, which is irrelevant.
- Going forward, the phrases machine learning and data science will dominate in popular lexicon.

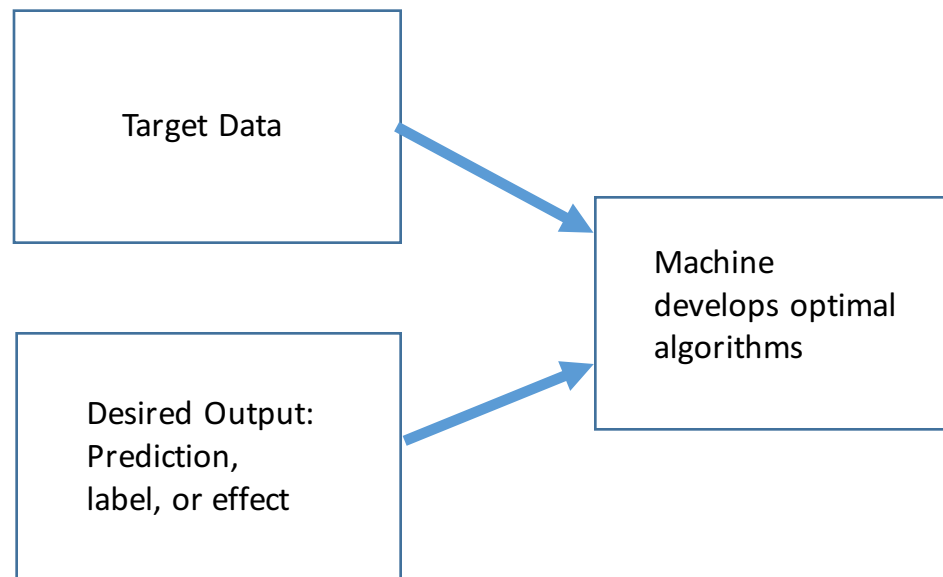
# Traditional ML Paradigm



# An Example

- This unlabeled image from the internet contains a four-legged hairy animal.
- 83.56% of *similar* images I have seen from the internet are labeled cats.
- Therefore, this image should be labeled as "cat".
- Entirely algorithmic.
- No artificial intelligence involved.

# Artificial Intelligence (AI) Paradigm



# Every ML Algorithm Has Three Components

1. Representation

2. Evaluation

3. Optimization

- Computer scientists applying ML are typically focused on classification and prediction.
- Consider the problem of predicting an incoming email to be "spam" given its contents (that is, conditional on its covariates).



## For Regression, It Is the Same

1. Representation:  $y_t = \rho y_{t-1} + x_t' \beta + \epsilon_t + \theta \epsilon_{t-1}$
  2. Evaluation: goodness of fit measures such as  $R^2$  or MSE.
  3. Optimization:  $\min SSR(\beta) \rightarrow (X'X)^{-1}X'y$  or maximum likelihood
- It should be noted that many popular ML techniques, such as neural networks, do not have global optima, but they work very well in practice.
  - Indeed, the ML mantra is often:
    - We have different tools for different problems.
    - How well do they work in practice?
    - How well do they scale (efficiency)?

# Useful ML Terms

- Features (independent variable, covariates, right-hand-side variables)
- Labels (dependent variable)
- Structured data (observation  $i$  at time  $t$  in a spreadsheet)
- Unstructured data (text)
- Training data
- Test (or hold out) data
- Supervised learning:  $y = X\beta + \epsilon$ 
  - Classification (classes of labels: red, blue, green)
  - Regression (continuous labels measuring some real world outcome)

## Useful ML Terms (cont'd)

- Unsupervised learning:  $y = f(x, \epsilon)$  or k-means clustering.
- Parallelization: why Google won the search engine war with distributed storage and distributed processing.
- UI: user interface, a type of HTML- or Java-based interface that allows users to interact with underlying hardware and software. (Jupyter notebook is an example.)
- API: application program interface, which assists analysts to input data from disparate sources, such as the Fed.
- IDE: integrated development environment, of which Jupyter notebook is a part.

# Notable Shortcomings

- Terms one does not typically hear at ML conferences.
  - Standard error.
  - p-value.
  - Statistical significance.
- Any discussion of causation or of "treatment effects" in the Angrist-Imbens-Rubin sense of evaluating the impact of an intervention.
- Computer scientists do not see their regression functions as measures of partial derivative effects.
- While these may be very interesting policy questions, these are not common questions in a ML textbook:
  - What is the effect of raising the minimum wage on employment levels?
  - What is the effect of building a hospital in this village on health outcomes?

# Economists Are Now Collaborating With Computer Scientists

- I presented a poster on treatment effects using big data and ML techniques at a recent ML conference.
- Susan Athey and Guido Imbens have some forthcoming papers on the use of ML in economics.
- DARPA's Next Generation Social Science (NGS2).
- As I point out to my friends who are computer scientists, economists have a long history of dealing with non-experimental data.
  - Heckman's inverse mills ratio as an omitted feature in regression to capture the selection effect.

# Economists Were Already at the Table

- Long tradition of large-scale time series modeling of markets and economies.
- Hal White on penalized regression and the LASSO in the early 1990's.
- George Tauchen and Ron Gallant on neural networks in the mid-1990's.
- More recent innovations in econometrics, however, have focused on natural experiments, RDD, and RCT.
- In my opinion, the use of a large suite of ML approaches will grow in econometrics (and I believe quite quickly).

# Let's Do Some of This Stuff

- Popular classification algorithms in Python.
- Simple neural networks in R.

# Classification



# Neural Networks

Comments and Questions