Big Data for Public Policy

Statistical Learning [Part 1]

Malka Guillot

ETH Zürich | 860-0033-00L

Prologue

References

- JWHT chap 1. & 2.1
- Kleinberg, Ludwig, Mullainathan, and Obermeyer (2015), "Prediction Policy Problems." American Economic Review, 105 (5), pp. 491-95.
- Mullainathan and Spiess (2017), "Machine Learning: An Applied Econometric Approach", Journal of Economic Perspectives, 31 (2), pp. 87-106,

Context

Today

- What is statistical learning?
- Statistics in social science – causality.
- Statistics in machine learning – prediction.
- Accuracy v. interpretability.

Next week

- Model accuracy.
- The bias-variance tradeoff.
- Classification

Table of contents

What is statistical learning?

Setting

- Input variables $\mathcal{X}$
  - AKA features, independent variables, predictors
- Output variables $\mathcal{Y}$
  - AKA dependent variables, outcomes, etc.

Statistical learning theory

$$f : \mathcal{X} \rightarrow \mathcal{Y}$$

$$\mathcal{X} \in \mathbb{R}^{n \times p}, \mathcal{Y} \in \mathbb{R}^{p}$$

SL= approaches for finding a function that accurately maps the inputs $\mathcal{X}$ to outputs $\mathcal{Y}$

Statistical model

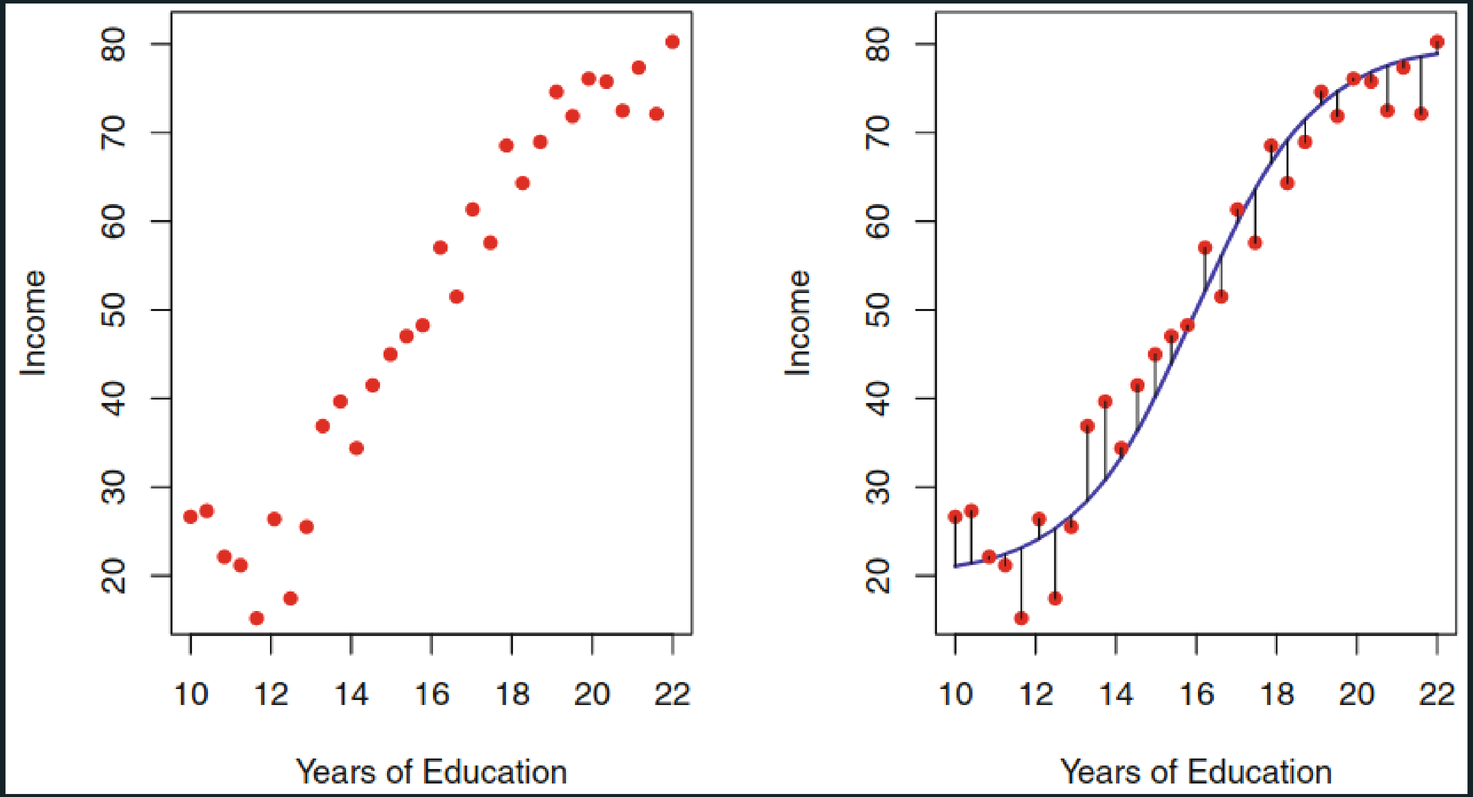
Concretely, finding $f(.)$ s.t.

$$Y = f(X) + \epsilon$$

- $f(X)$ is an unknown function of a matrix of predictors $X = (X_1, \cdots, X_p)$,
- Y: a scalar outcome variable
- an error term ϵ with mean zero.
- While X and Y are known, $f(\cdot)$ is unknown.

Goal of statistical learning: to utilize a set of approaches to estimate the “best” $f(\cdot)$ for the problem at hand.

Example: income as a function of education



Prediction

- Predict Y by $\hat{Y} = \hat{f}(X)$
- When do we care about "pure prediction"?
  - X readily available but Y is not
- $\hat{f}$ can be a **block box**:
  - the only concern is accuracy of the prediction

Inference

- Understanding the way that Y is affected as $X_1, \ldots, X_p$ change
  - Which predictors are associated with the response?
  - What is the relationship between the response and each predictor?

$\Rightarrow$ $\hat{f}$ is cannot be a **black box** anymore

Approach in social science

- Objective: Understanding the way that Y is affected as $X_1, \ldots, X_p$ change
- The goal not necessarily to make predictions for Y
- Often linear function to estimate Y: $f(X) = \sum_{i=1}^{p} \beta_i x_i$
- Assume $\epsilon \sim N(0, \sigma^2)$
- Parameters β are estimated by minimizing the sum of squared errors

$$Y = \sum^{p} \beta_i x_i + \epsilon$$

Approach in social science: causality

$$Y = \beta_0 + \beta_1 T + \sum_{i=1}^{p-1} \beta_i x_i + \epsilon$$

- Interested in the values of one or two parameters and whether they are **causal** or not.
- Framework to interpret statistical causality: **Rubin (1974)**
- β_1 measures the extent to which ΔX_t will affect ΔY_{t+1}

Approach in social science: causality

- Causal inference requires that $T \perp \epsilon$ or $T|X \perp \epsilon$

$\rightarrow$ can be achieved through randomization of T

- This implies that we are not really all that interested in choosing an optimal $f(.)$
- (We want to estimate unbiased coefficients)

Approach in machine learning: prediction

$$\hat{Y} = \hat{f}(X)$$

- Objectives:
  - find the "best" $f(\cdot)$ and the "best" set of X's which give the best predictions, $\hat{Y}$
  - **Accuracy**: find the function that **minimize the difference between *predicted* and *observed* values**
  - (We want to minimize prediction error)

Reducible and irreducible error

$\hat{f}(X) = \hat{Y}$ estimated function

$f(X) + \epsilon = \hat{Y}$ true function

- **Reducible error**: $\hat{f}$ is used to estimate f, but not perfect $\rightarrow$ accuracy can be improved by adding more features
- **Irreducible error**: ϵ = all other features that can be used to predict f $\rightarrow$ unobserved $\rightarrow$ irreducible

Reducible and irreducible error

$$E(Y - \hat{Y})^2 = E[f(X) + \epsilon - \hat{f}(X)]^2$$

$$= \underbrace{[f(X) - \hat{f}(X)]^2}_{Reducible} + \overbrace{Var(\epsilon)}^{Irreducible}$$

$\Rightarrow$ **Objective:** estimating f with the aim of minimizing the reducible error

How do we estimate f?

Context

We use observations to "teach" our ML algorithm to predict outcomes

- **Training data**: $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ where $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$
- Goal: use the training data to estimate the unknown function f
- 2 types of SL methods: **parameteric vs. nonparametric**

Parametric methods

Model-based approaches, 2 steps:

1. Specify a ***parametric* (functional) form** for $f(X)$, e.g. linear:

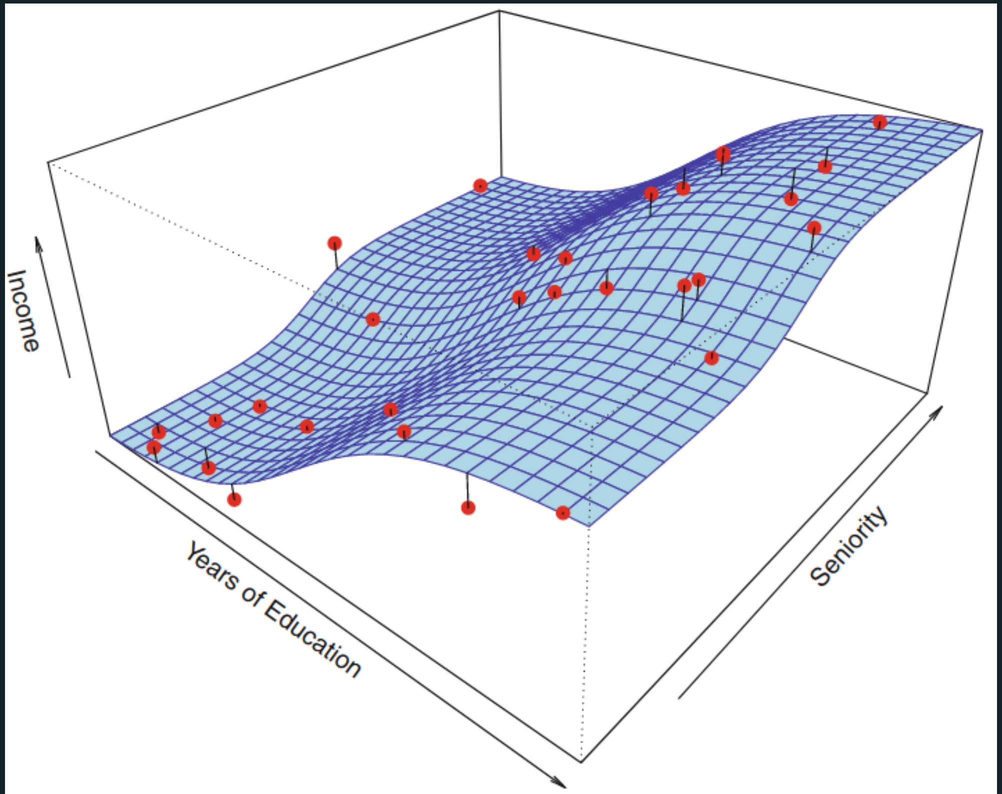
$$f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

   (Parametric means that the function depends on a finitenumber of parameters, here $p + 1$).

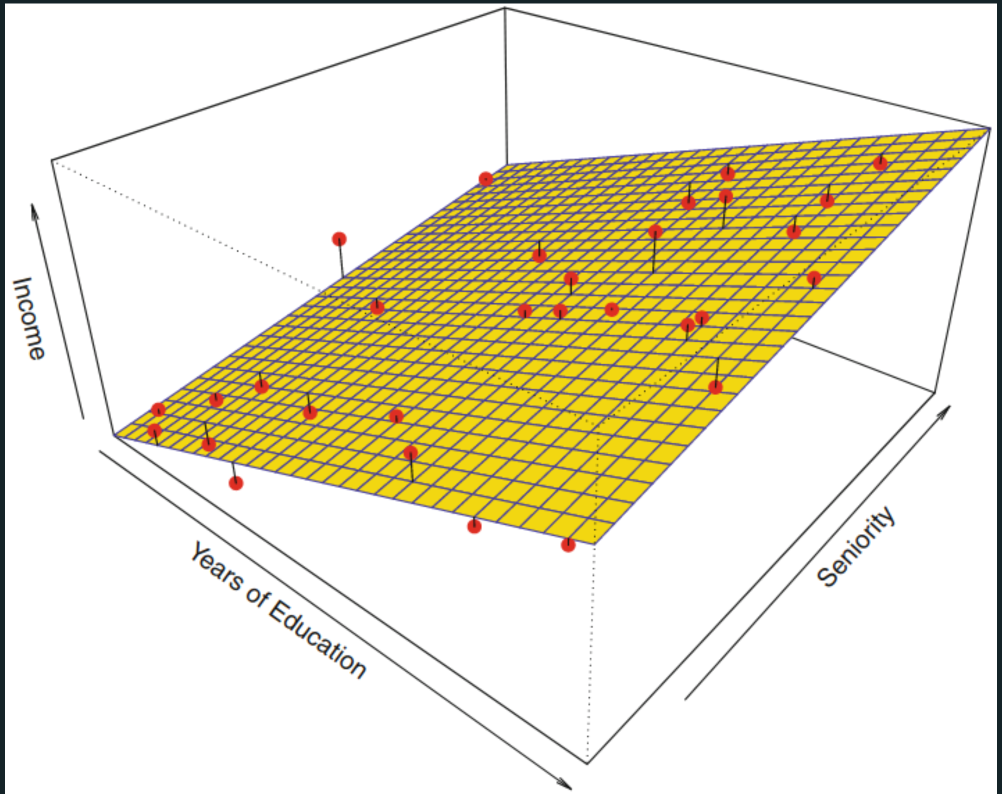
2. **Training**: Estimate the parameters by OLS and predict Y by

$$\hat{Y} = \hat{f}(X) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p$$

True function



Linear estimate



Parametric methods -- issues

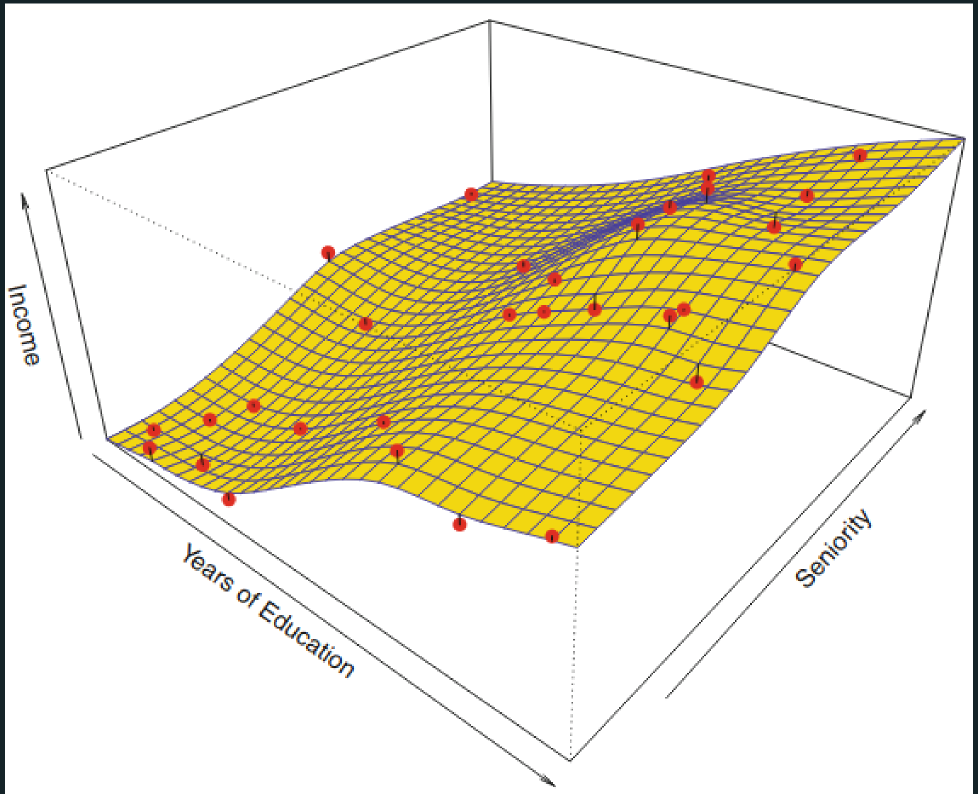
Misspecification of $f(X)$

1. Rigid models (e.g. strictly linear) may not fit the data well
2. More flexible models require more parameter estimation $\rightarrow$ **overfitting**

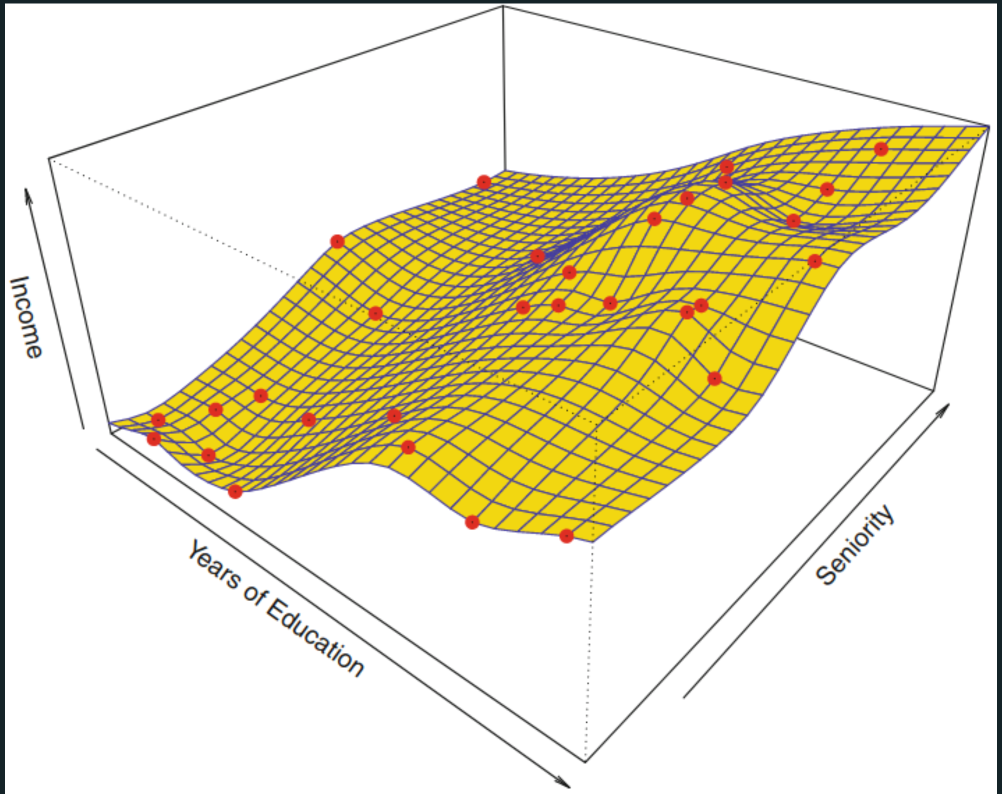
Non-parametric methods

- **No assumptions** about the functional form of f
- Estimates a function only **based on the data itself**.
- **Disadvantage**: very large number of observations is required to obtain an accurate estimate of f

“Smooth” nonlinear estimate



Rough nonlinear estimate with perfect fit $\Rightarrow$ overfit



Recap: parametric vs. non-parametric approaches

Which of the following applies to parametric methods?

- [ ] Only estimating a set of parameters
- [ ] Gives insight on the data when nothing is known
- [ ] Better predictions with little data
- [ ] Rely on model assumptions

Accuracy and interpretability tradeoffs

- **More accurate** models often require estimating **more parameters** and/or having more flexible models
- Models that are better at prediction generally are **less interpretable**.
- For inference, we care about interpretability.

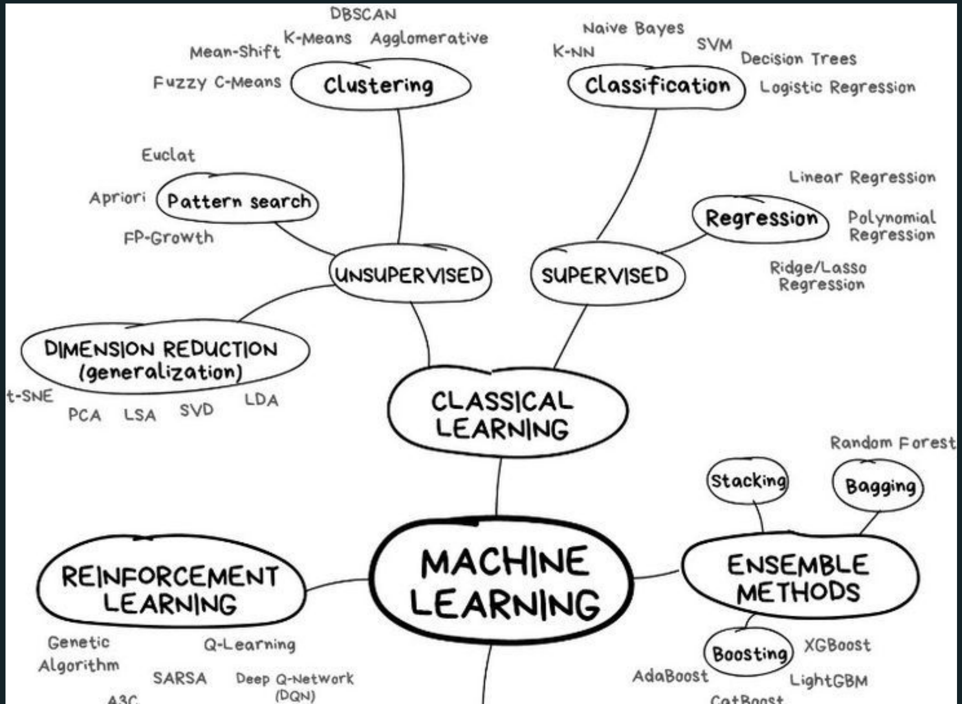
$\rightarrow$ More on this next week!

Machine Learning: overview and examples

Supervised vs. unsupervised learning

- **Supervised learning**: estimating functions with known observation and outcome data.
  - We observe data on Y and X and want to learn the mapping $\hat{Y} = \hat{f}(X)$
  - **Classification** when $\hat{Y}$ discrete; **regression** when $\hat{Y}$ continuous
- **Unsupervised learning**: estimating functions without the aid of outcome data.

The Machine learning landscape



Examples: Studies using ML for p rediction

- Glaeser, Kominers, Luca, and Naik (2016) use images from Google Street View to measure block-level income in New York City and Boston
- Jean et al. (2016) train a neural net to predict local economic outcomes from satellite data in African countries
- Chandler, Levitt, and List (2011) predict shootings among high-risk youth so that mentoring interventions can be appropriately targeted
- Kleinberg, Lakkaraju, Leskovec, Ludwig, and Mullainathan (2018) predict the crime probability of defendants released from investigative custody to improve judge decisions
- Kang, Kuznetsova, Luca, and Choi (2013) use restaurant reviews on Yelp.com to predict the outcome of hygiene inspections
- Huber and Imhof (2018) use machine learning to detect bid- rigging cartels in

The Machine learning workflow

1. Look at the big picture.
2. Get the data.
3. Discover and visualize the data to gain insights.
4. Prepare the data for Machine Learning algorithms.
5. Select a model and train it.
6. Fine-tune your model.
7. Present your solution.8. Launch, monitor, and maintain your system

Aurelien Geron, *Hands-on machine learning with Scikit-Learn & TensorFlow*, Chapter 2

Conclusion:

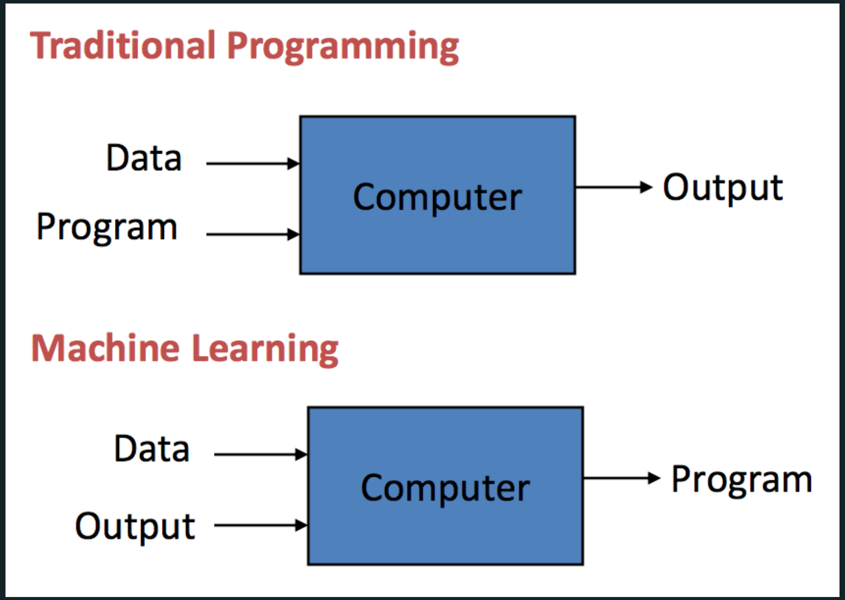
Econometrics vs. Machine Learning

Econometrics vs. Machine Learning (1)

- **Common objective**: to build a predictive model, for a variable of interest, using explanatory variables (or features)
- **Different cultures**:
  - *E*: probabilistic models designed to describe economic phenomena
  - *ML*: algorithms capable of learning from their mistakes

Charpentier A., Flachaire, E. & Ly, A. (2018). Econometrics and Machine Learning. *Economics and Statistics*, 505-506, 147–169.

Econometrics vs. Machine Learning (2)



- **Classical computer programming:** humans input the rules and the

Researcher vs. policy analysist

- The frontier can be thin
- I will sometimes be speaking from the point of view of an economist, but:
  - The model-based vs. algorithm-based problematics transfers to other social sciences
  - I try to cover a wide range of topics in the literature
  - You are welcome to propose relevant papers
- All aim at *using data to solve problems*