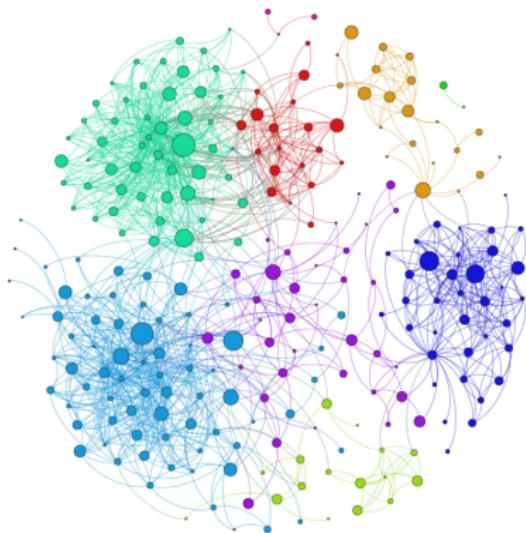


# Data Analysis for Economics

## à Modern Introduction

Jiaming Mao

Xiamen University



Copyright © 2017–2019, by Jiaming Mao

This version: Spring 2019

Contact: [jmao@xmu.edu.cn](mailto:jmao@xmu.edu.cn)

Course homepage: [jiamingmao.github.io/data-analysis](https://jiamingmao.github.io/data-analysis)



All materials are licensed under the [Creative Commons Attribution-NonCommercial 4.0 International License](#).

# Data Analysis

*“What’s in a name? that which we call a rose,  
By any other name would smell as sweet.” – Juliet*

Machine Learning → Statistics → Econometrics

- Along this spectrum, the focus moves from prediction and pattern discovery to inference about causality and the underlying mechanisms that generate the observed data.

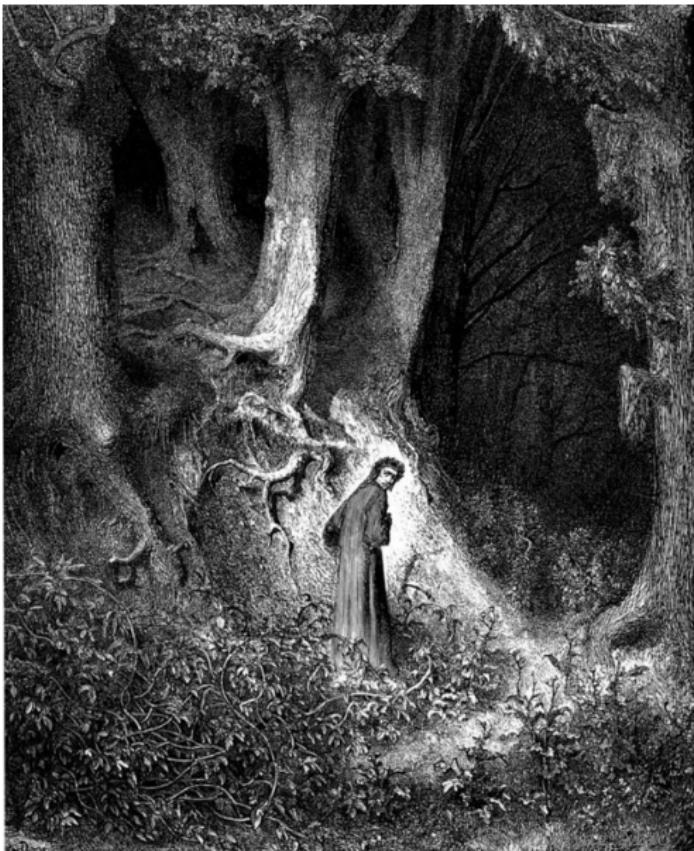
# Road Map

1 Statistical Learning

2 Causal Inference

# Road Map

Thematically, we follow the journey of a hero determined to seek knowledge from data, who departs the *forest of ignorance*,



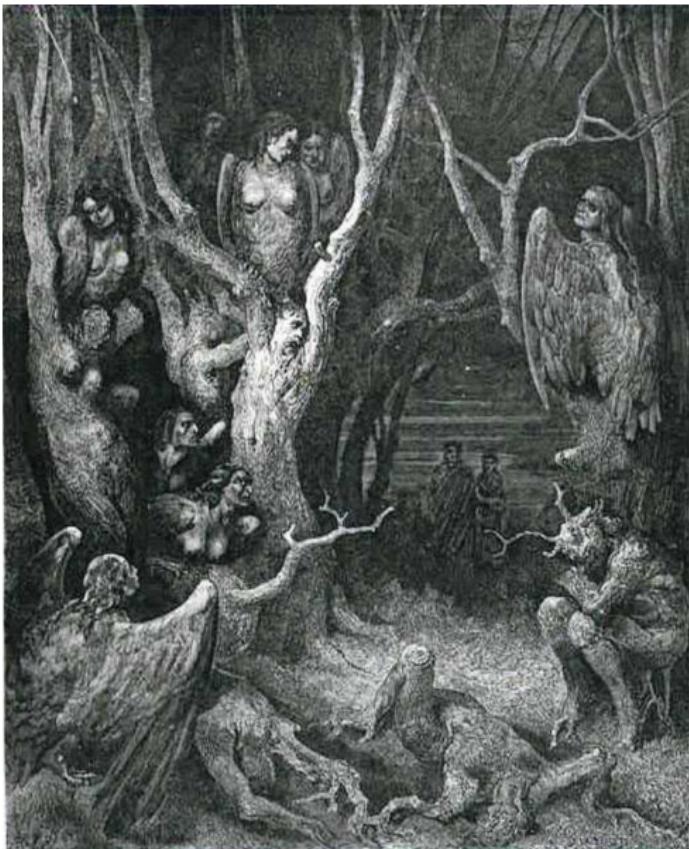
# Road Map

... and journeys to the *realm of patterns*, where patterns in data are discovered and used to make predictions,



## Road Map

... along the way he encounters the false prophets of *correlation equals causation*,



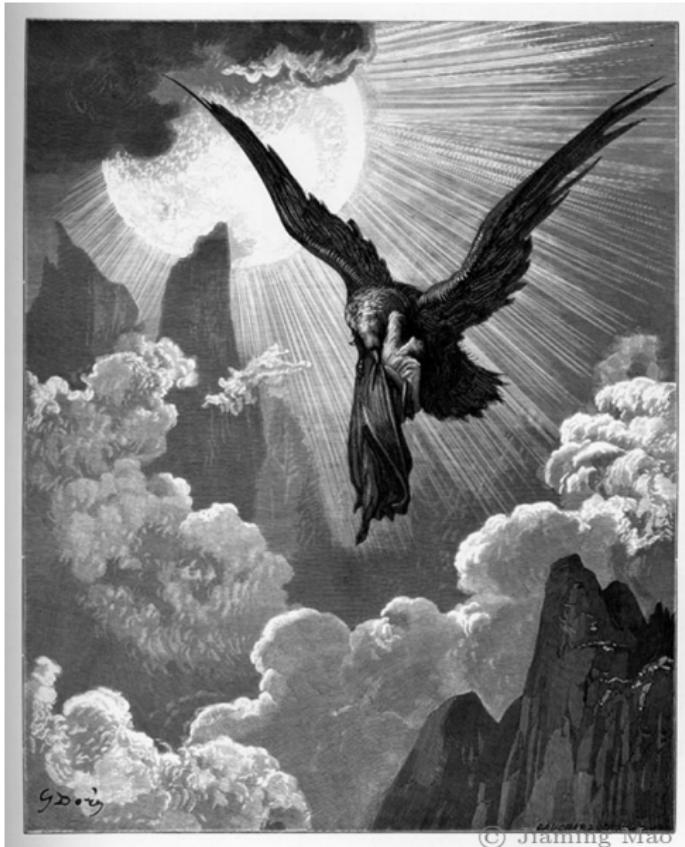
# Road Map

... and then arrives at the *land of causality*, where people are serious about whether any two sets of observed phenomena are linked causally,



# Road Map

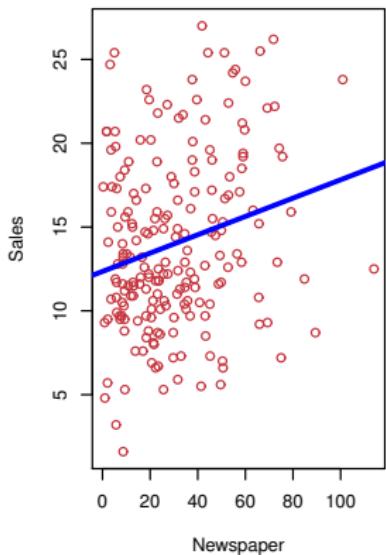
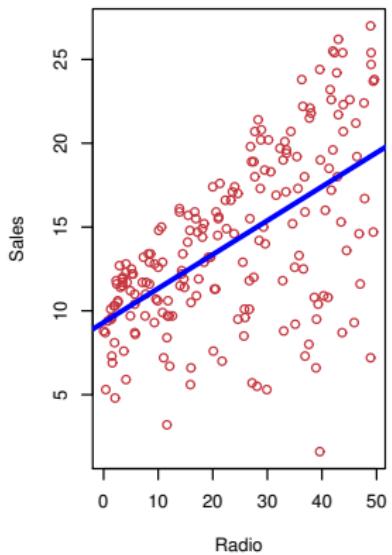
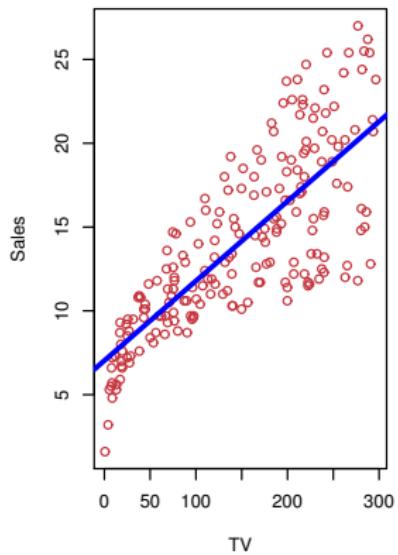
... from where our hero finally reaches the *mount of scientific discovery*, where the mechanisms that generate the observed phenomena are investigated in the hope of attaining true knowledge about the world.



# Statistical Learning

- Given variables  $x$  and  $y$ , how do we characterize the statistical relationship between the two?
  - $p(x, y)$  : joint distribution of  $x$  and  $y$
- Oftentimes, we may not be interested in characterizing the full joint distribution  $p(x, y)$ . Instead, we are interested in predicting the value of  $y$  based on observed  $x$ .
  - We want to find a function  $f(x)$  for predicting  $y$  given values of  $x$ .

# Statistical Learning



Advertising and Sales

# Statistical Learning

Let

$$y = f(x) + e$$

, where  $e$  is an error term.

What is the function  $f$  that produces the **best** prediction of  $y$  given  $x$ ?

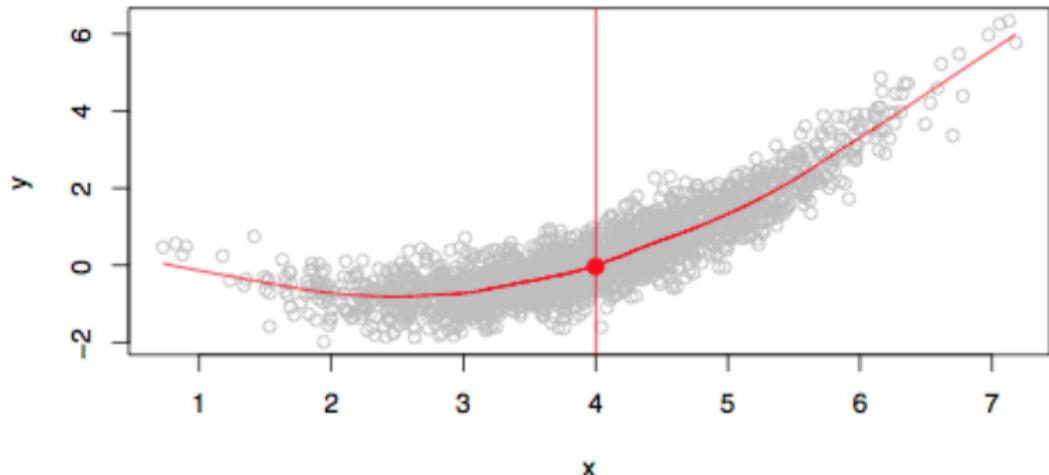
- Depends on how we measure “best.” Common choice: minimizing the expected L2 loss<sup>1</sup>  $E[(y - f(x))^2] \Rightarrow f(x) = E[y|x]$ .
- $f(x) = E[y|x]$  is the **target function** that we want to learn<sup>2</sup>.

---

<sup>1</sup>Also commonly called the **mean squared error (MSE)**.

<sup>2</sup>**Learning** is also called **estimation**. We will use the two terms interchangeably.

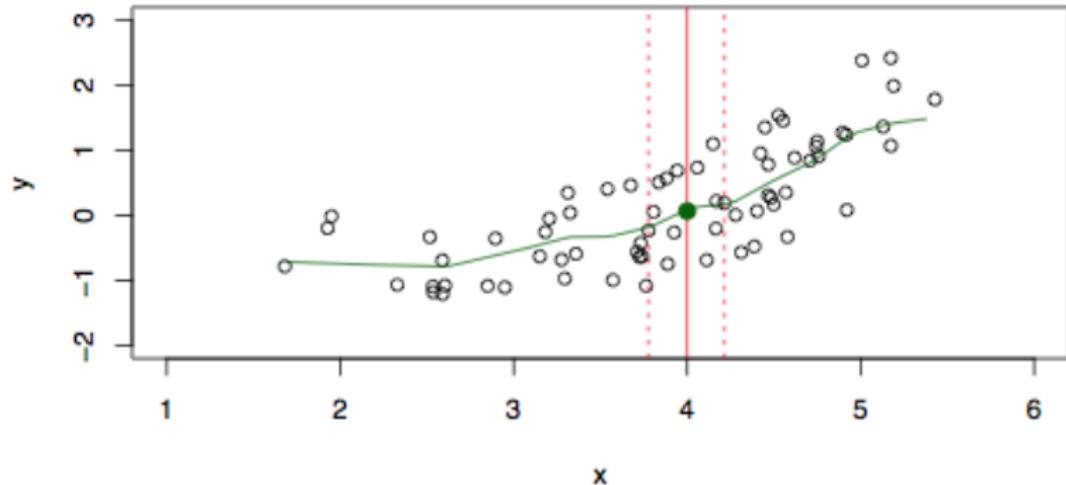
# Learning f



$$\hat{f}(x = 4) = \text{Ave}(y|x = 4)$$

## Learning f

- Typically we have few if any data points at a specific value of  $x$ .
- One solution: relax the set of  $x$  over which  $y$  is averaged.



$$\hat{f}(x = 4) = \text{Ave}(y | x \in \mathcal{N}(x = 4))$$

, where  $\mathcal{N}(x)$  is some neighborhood of  $x$ .

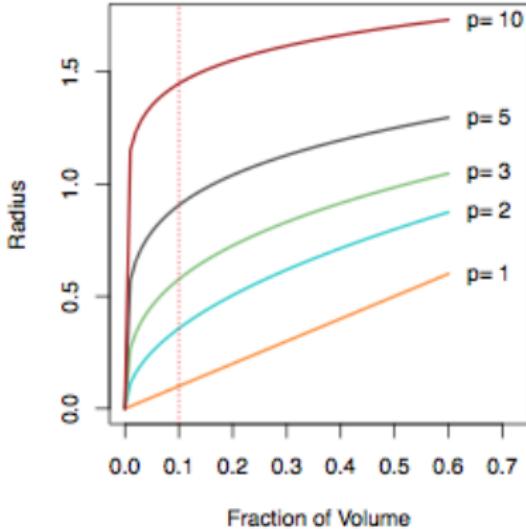
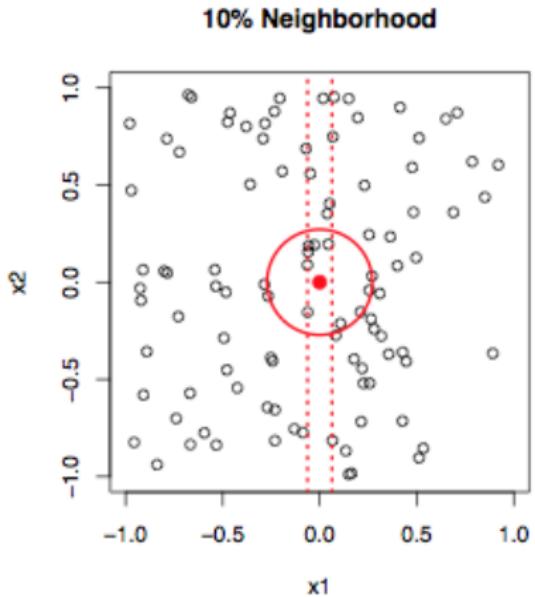
## Learning f

- When  $x$  is multi-dimensional, i.e.  $x = (x_1, \dots, x_p)$ , nearest neighbor averaging can work well for small  $p$  and large  $N^3$ .
- Nearest neighbor methods can be lousy when  $p$  is large, because neighbors tend to be far away in high dimensions.
  - ▶ This is called the **curse of dimensionality**.

---

<sup>3</sup> $N$  : the number of data points, i.e.  $N$  is the size of the training data.

# Learning f

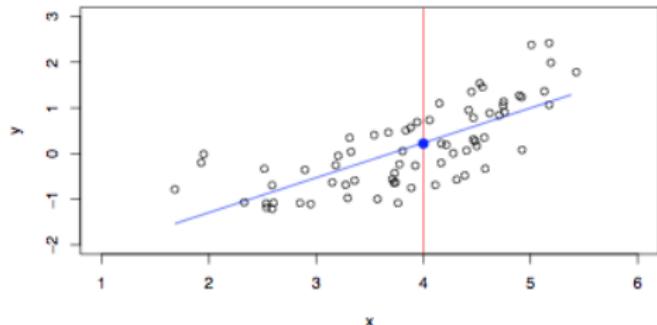


Nearest neighbor and the curse of dimensionality

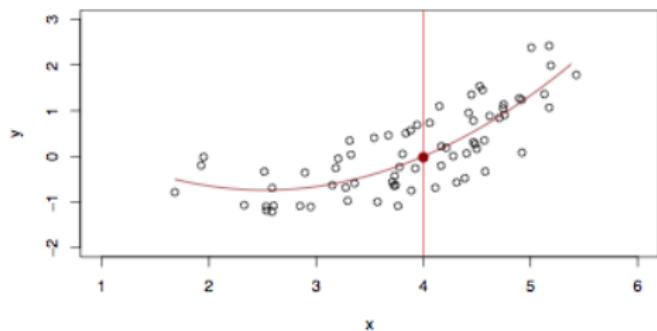
# Learning f

- **Parametric methods** of estimating  $f(x)$  assume a specific functional form.
  - ▶ Linear regression:  $\hat{f}(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p = \beta' x$
- **Nonparametric methods** do not make explicit assumptions about the functional form of  $f(x)$ .
  - ▶ Nearest neighbor averaging is a nonparametric method.

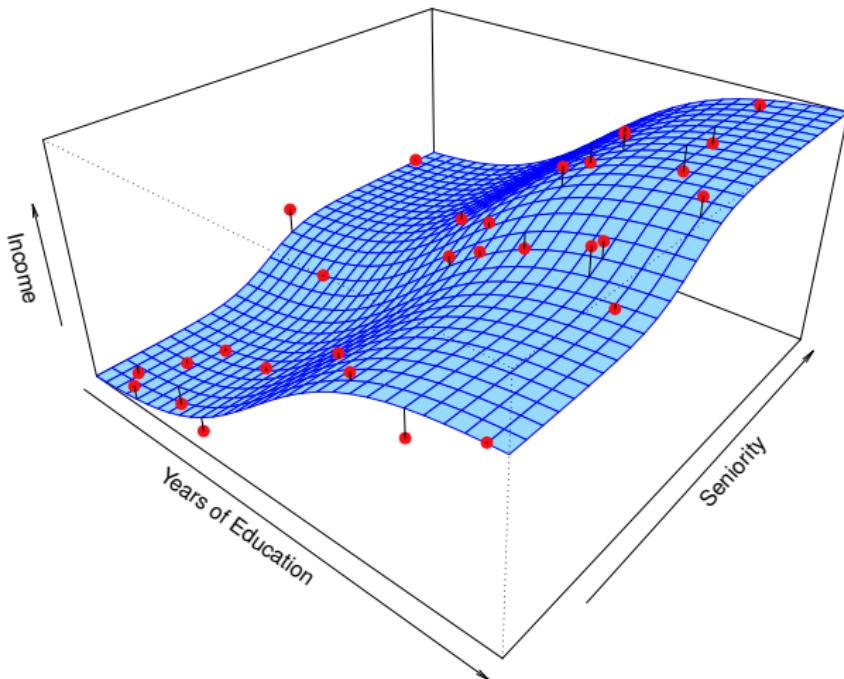
A linear model  $\hat{f}_L(X) = \hat{\beta}_0 + \hat{\beta}_1 X$  gives a reasonable fit here



A quadratic model  $\hat{f}_Q(X) = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 X^2$  fits slightly better.

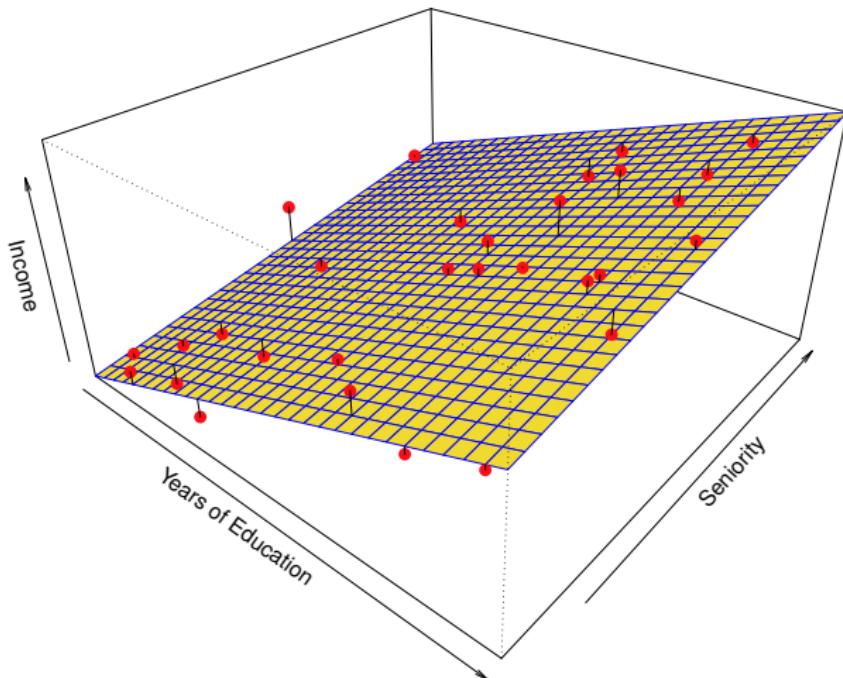


# Learning f



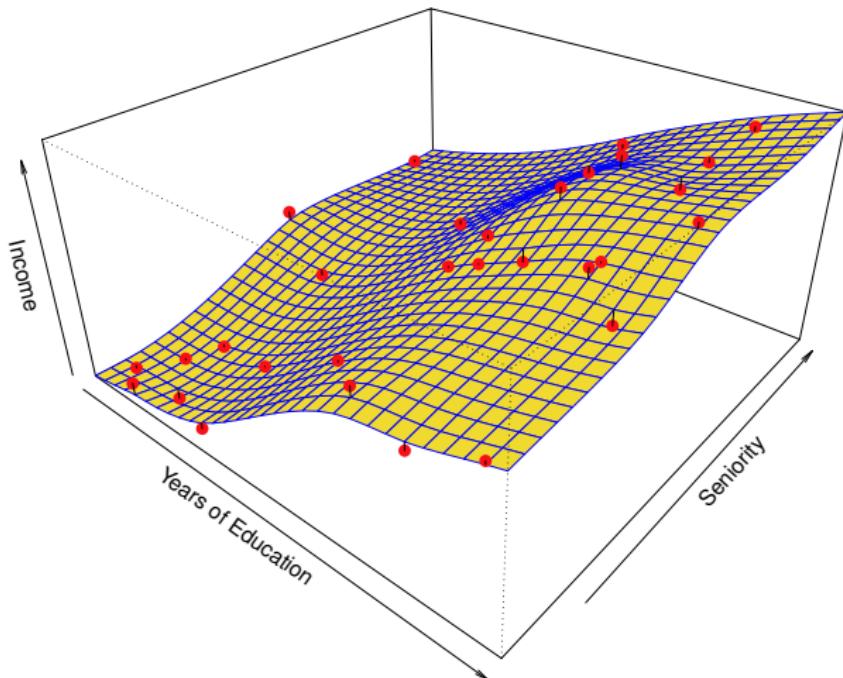
Age, Education and Income: True  $f$

# Learning f



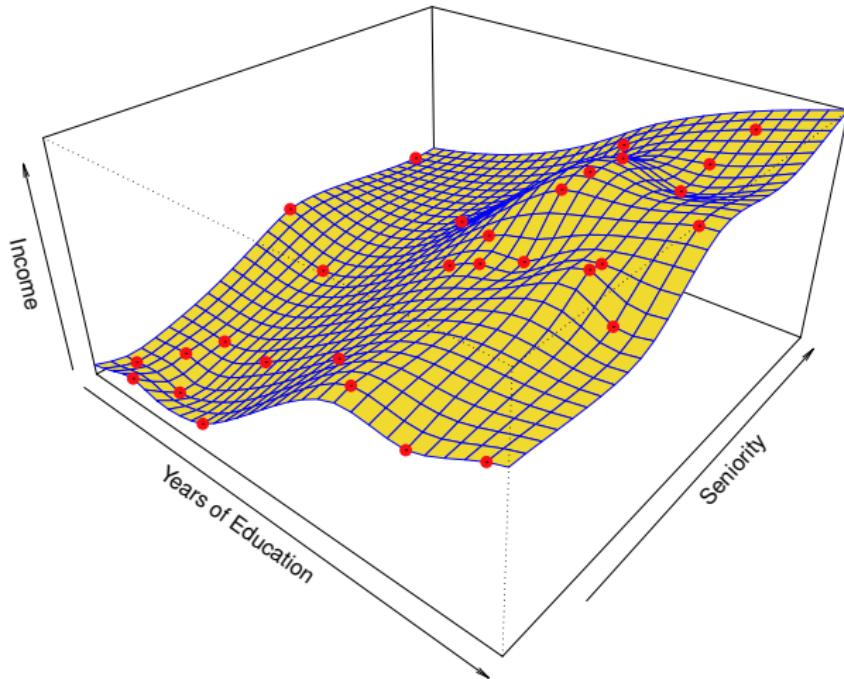
Age, Education and Income: Linear Fit

# Learning f



Age, Education and Income: Thin-plate Spline Fit (Smooth)

# Learning f



Age, Education and Income: Thin-plate Spline Fit (Rough)  
Here  $\hat{f}$  fits the data perfectly:  $\hat{f}(x)$  contains not only  $f(x)$  but also  $e$ .  
This is a case of *overfitting*.

## Assessing the Goodness of Fit

- Let  $\mathcal{D}_{TR} = \{(x_1, y_1), \dots, (x_N, y_N)\}$  denote the data on which we estimate  $f$ . This is called **training data**.
- We can assess how well  $\hat{f}$  fits the training data by calculating the **training error**:

$$ERR_{TR} = \frac{1}{N} \sum_{i \in \mathcal{D}_{TR}} (y_i - \hat{f}(x_i))^2$$

- However, what we are really interested in is how well  $\hat{f}$  predicts previously unseen data.

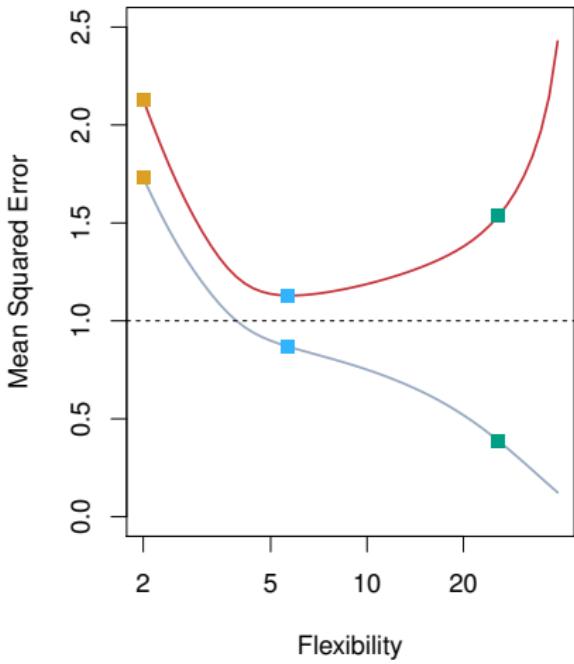
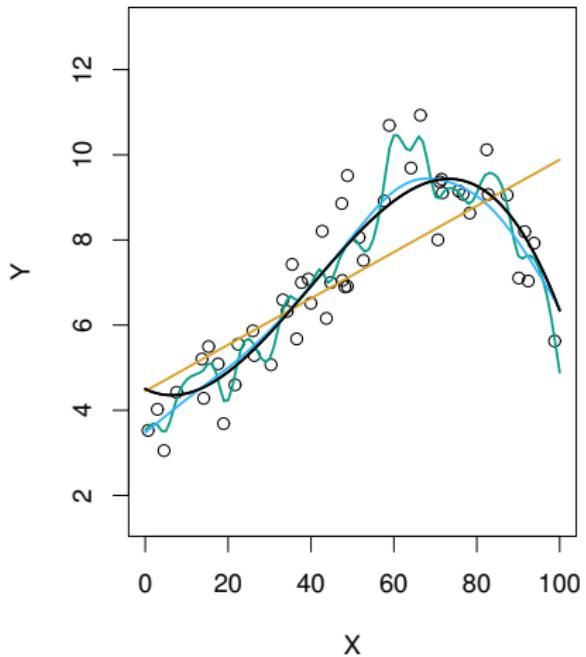
## Assessing the Goodness of Fit

- To this end, we can apply  $\hat{f}$  to a set of **test data**,  $\mathcal{D}_{TE} = \{(x_1, y_1), \dots, (x_M, y_M)\}$ , and calculate the **test error**:

$$ERR_{TE} = \frac{1}{M} \sum_{i \in \mathcal{D}_{TE}} (y_i - \hat{f}(x_i))^2$$

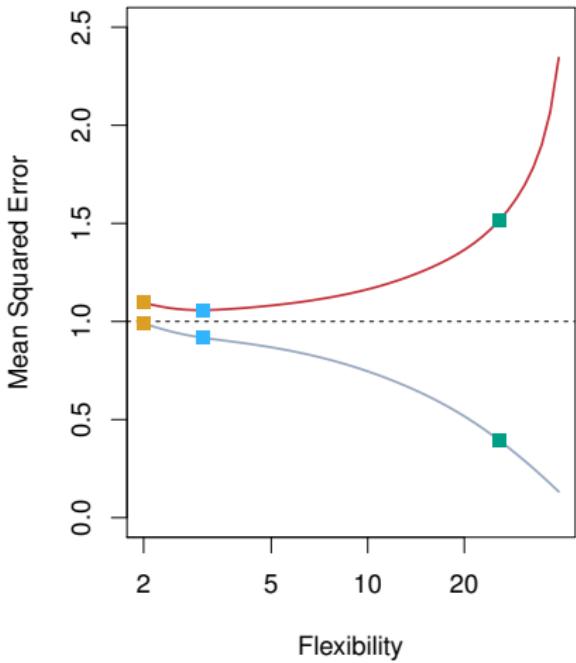
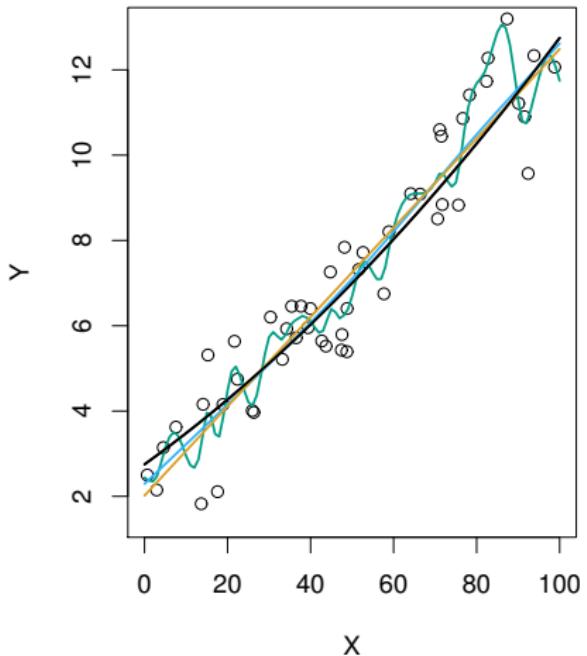
- When  $M \rightarrow \infty$ ,  $ERR_{TE} \rightarrow \underbrace{E \left[ (y - \hat{f}(x))^2 \right]}_{\text{prediction error}}$ .

# Assessing the Goodness of Fit



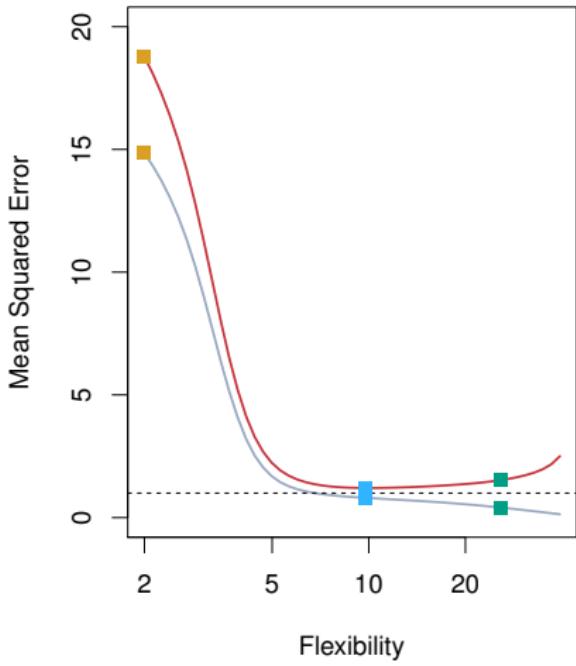
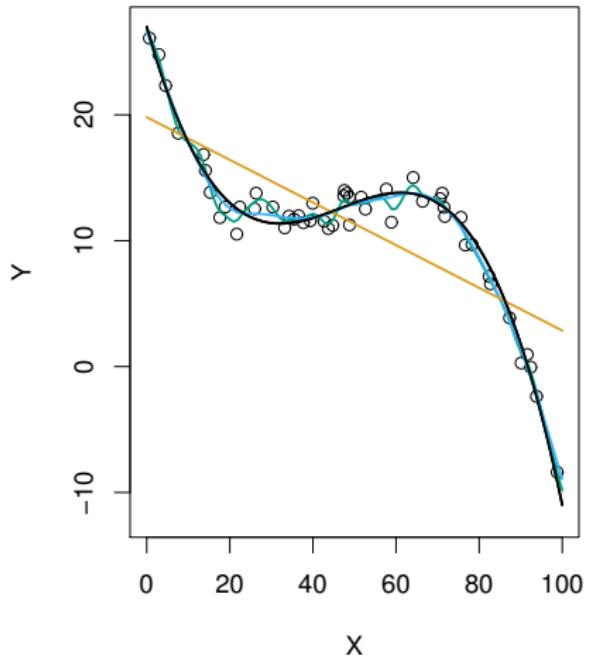
Left: true  $f$  (black), linear fit (orange), smoothing spline fits (blue & green).  
Right: training error (grey), prediction error (red),  $\text{Var}(e)$  (dashed).

# Assessing the Goodness of Fit



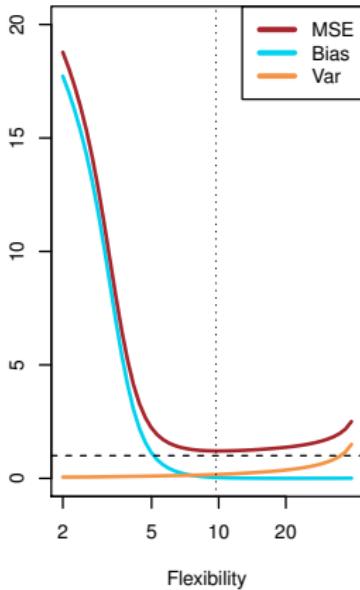
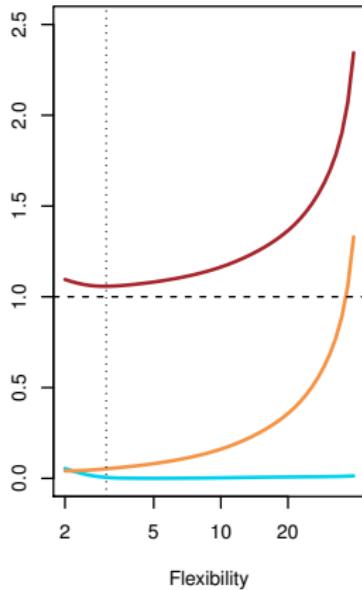
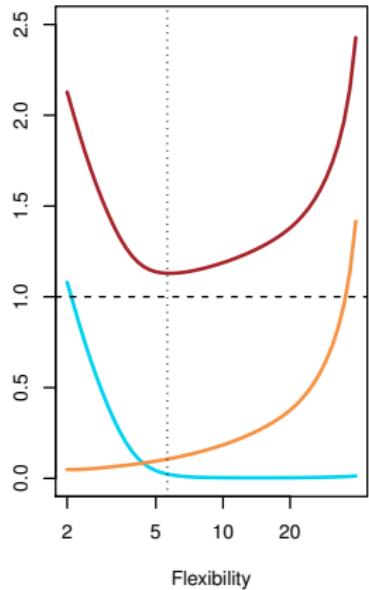
Left: true  $f$  (black), linear fit (orange), smoothing spline fits (blue & green).  
Right: training error (grey), prediction error (red),  $\text{Var}(e)$  (dashed).

# Assessing the Goodness of Fit



Left: true  $f$  (black), linear fit (orange), smoothing spline fits (blue & green).  
Right: training error (grey), prediction error (red),  $\text{Var}(e)$  (dashed).

# Assessing the Goodness of Fit



Bias-variance trade-off for the three examples

# The Bias-Variance Trade-off

When  $f(x) = E(y|x)$ , for any estimate  $\hat{f}$  of  $f$  and  $\hat{y} = \hat{f}(x)$ ,

$$\begin{aligned}E[(y - \hat{y})^2] &= E\left[\left(f(x) - \hat{f}(x)\right)^2\right] + \underbrace{\text{Var}(e)}_{\text{Irreducible}} \\&= \text{Var}(\hat{f}(x)) + [\text{bias}(\hat{f}(x))]^2 + \text{Var}(e)\end{aligned}$$

, where  $\text{bias}(\hat{f}(x)) \equiv E[f(x) - \hat{f}(x)]$ .

# The Bias-Variance Trade-off

- $\text{Var}(\hat{f})$  refers to the amount by which  $\hat{f}$  would change if we estimated it using a different training data set.
- As a general rule, as we use more flexible methods, bias ( $\hat{f}$ ) will decrease and  $\text{Var}(\hat{f})$  will increase.
  - ▶ More flexible methods tend to have higher variance because they follow the data more closely. Thus changing any of the data points may cause the estimate  $\hat{f}$  to change considerably.

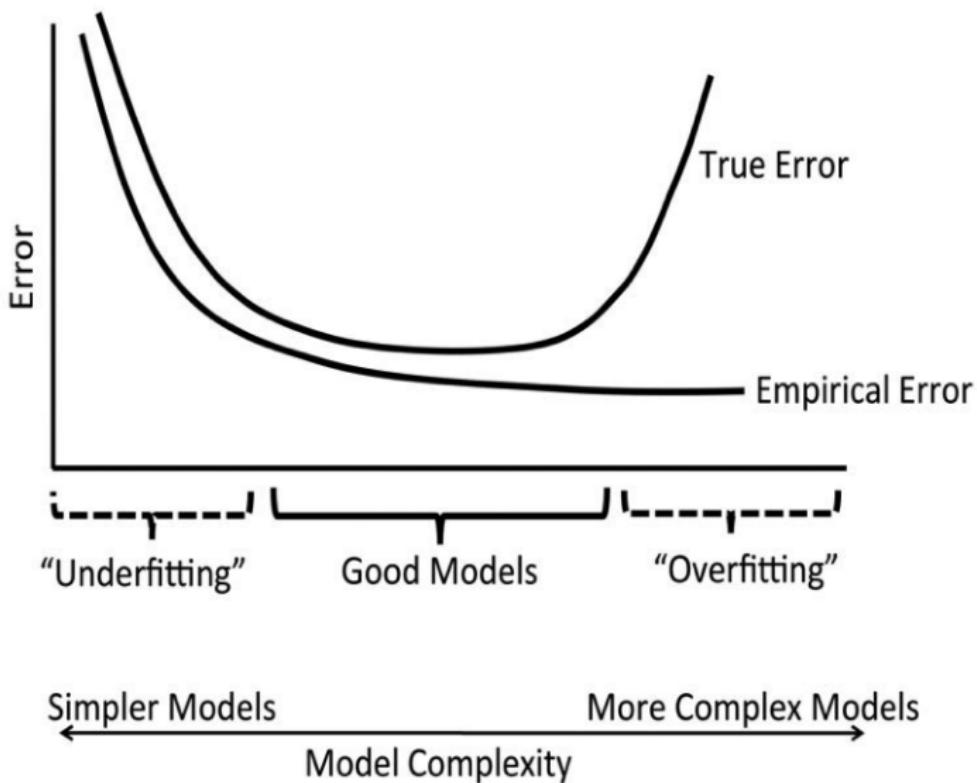
# The Bias-Variance Trade-off

- As the flexibility of the estimation method increases, we observe a monotone decrease in the training error and a U-shape in the prediction error.
- This is due to the **bias-variance trade-off**: as we increase the flexibility of the estimation methods, the bias tends to initially decrease faster than the variance increases. Then at some point increasing flexibility has little impact on the bias but starts to significantly increase the variance.

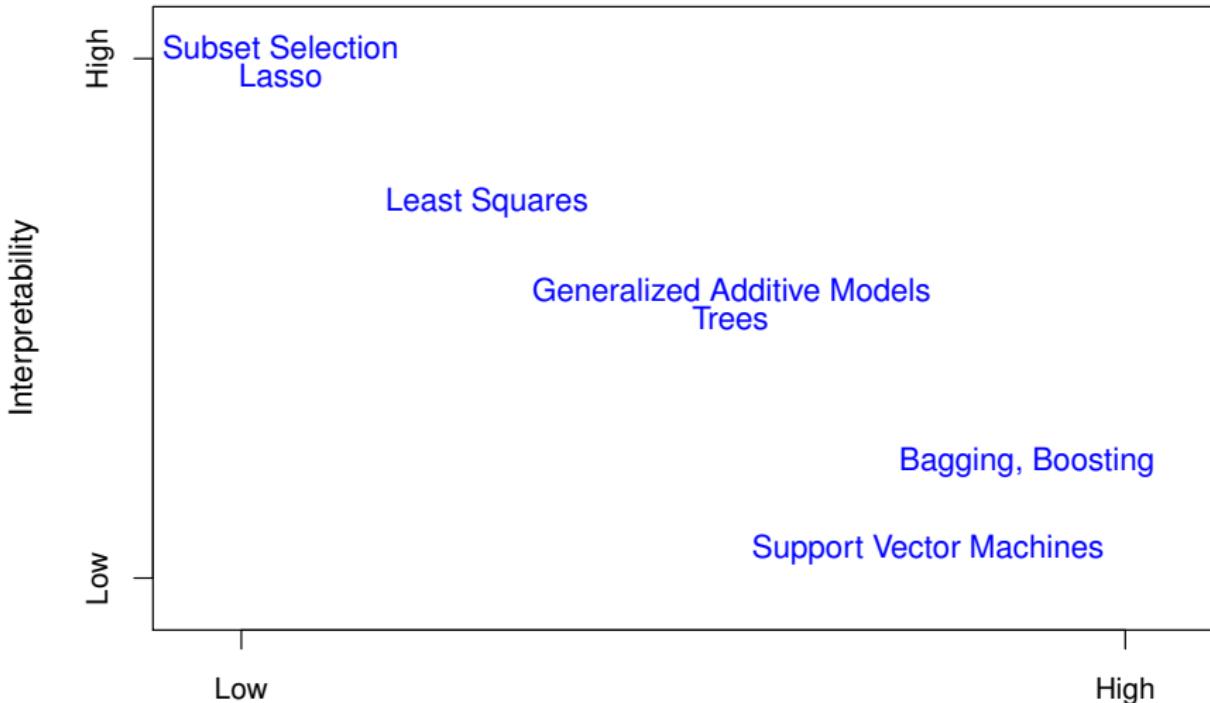
# The Bias-Variance Trade-off

- The bias-variance trade-off is a trade-off because it is easy to obtain a method with extremely low bias but high variance (e.g., by drawing a curve that passes through every single training observation) or a method with very low variance but high bias (by fitting a horizontal line to the data). The challenge lies in finding a method for which both the variance and the bias are low.
- **Overfitting** refers to the case in which a less flexible model would have yielded a smaller prediction error.

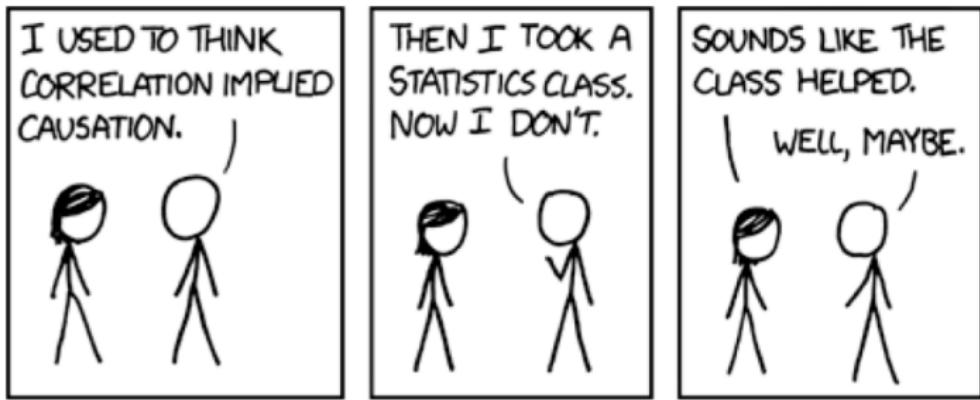
# The Bias-Variance Trade-off



# The Flexibility-Interpretability Trade-off



# Causal Inference

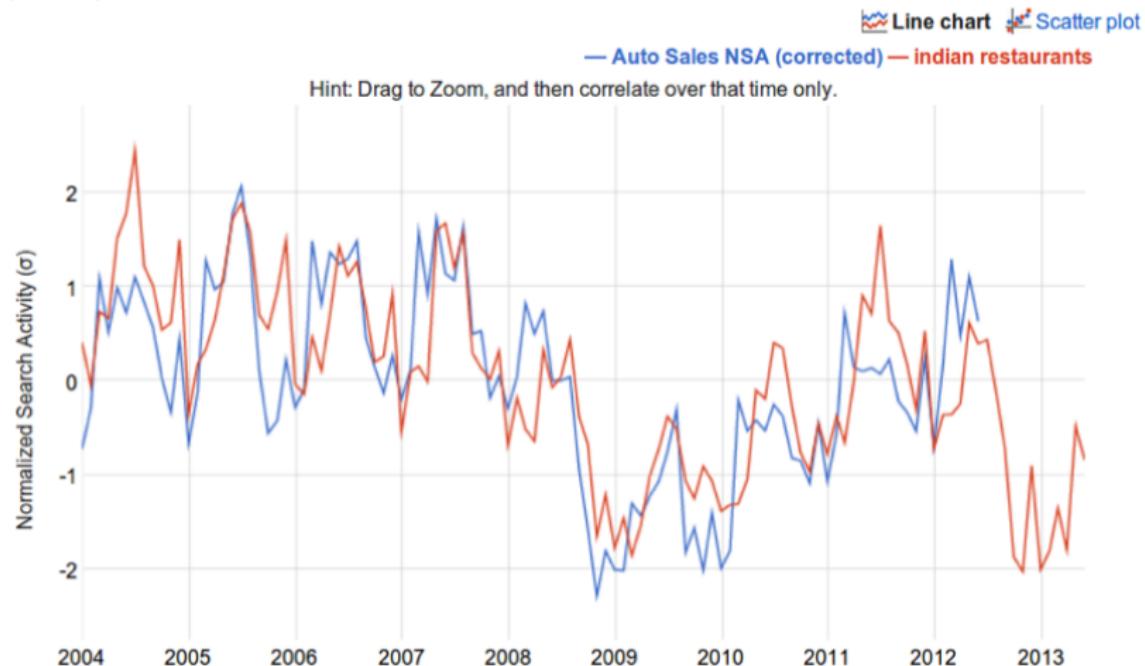


# Causal Inference

- Learning  $p(x, y)$  or  $E[y|x]$  tells us nothing about whether there exists a causal relationship between  $x$  and  $y$ .
- **Causal inference** is concerned with the following questions:
  - ① Does  $x$  have a causal effect on  $y$ ? If so, how large is the effect?  
**(causal effect learning)**
  - ② If a causal effect exists, what is the mechanism by which it occurs?  
**(causal mechanism learning)**

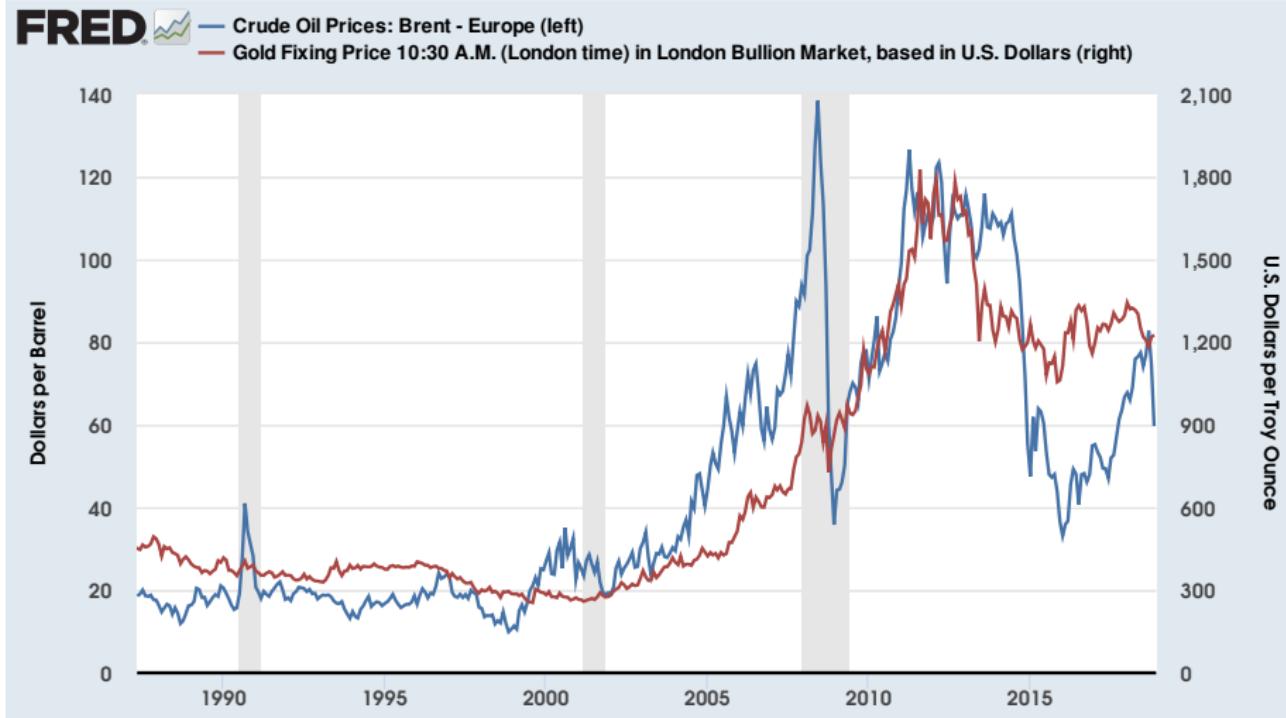
# Correlation does not imply Causation

User uploaded activity for Auto Sales NSA (corrected) and United States Web Search activity for indian restaurants  
 $(r=0.7848)$



Auto Sales and Search for Indian Restaurants. Source: [Google Correlate](#)

# Correlation does not imply Causation



## Russell's Chicken

Bertrand Russell<sup>a</sup> told the following cautionary tale of the perils of not understanding causal mechanisms:

*A chicken infers, on repeated evidence, that when the farmer comes in the morning, he feeds her. The inference serves her well until Christmas morning, when he wrings her neck and serves her for dinner.*

---

<sup>a</sup>Russell (1912), via Deaton and Cartwright (2018).

# Seeing vs. Doing

The do operator:

$$\text{do}(x = a) : \text{set } x = a$$

---

- Barometer readings are useful for predicting rain:

$$\Pr(\text{rain} \mid \text{barometer} = \text{low}) > \Pr(\text{rain} \mid \text{barometer} = \text{high})$$

- But hacking a barometer won't change the probability of raining:

$$\Pr(\text{rain} \mid \text{do}(\text{barometer} = \text{low})) = \Pr(\text{rain} \mid \text{do}(\text{barometer} = \text{high}))$$

# Seeing vs. Doing

- Doing: if  $x$  has a causal effect on  $y$ , then we can change  $x$  and expect it to cause a change in  $y$ .
- Seeing: If  $x$  is correlated<sup>4</sup> with  $y$  but does not have a causal effect on  $y$ , then we can only observe the correlation without the ability to change  $y$  by manipulating  $x$ .
- Holland (1986): “*No causation without manipulation.*”

---

<sup>4</sup>We use the term “correlation” in its broad sense to mean statistical dependence (association).

# Causal vs. Statistical Predictions

- **Causal prediction:** What will  $y$  be if I set  $x = a$ ?
  - ▶  $E[y|\text{do}(x = a)]^5$
- **Statistical prediction:** What will  $y$  be if I observe  $x = a$ ?
  - ▶  $E[y|x = a]$

---

<sup>5</sup>Assuming we minimize the expected L2 loss in prediction.

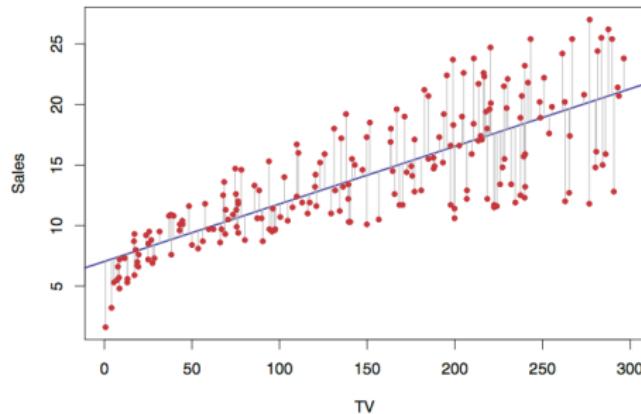
## Hospitalization and Health

Average health (assigning a 1 to poor health and a 5 to excellent health) contrasting those who have been an inpatient in the past 12 months and those who have not (tabulated from the 2005 NHIS):

Group	Sample Size	Mean health status	Std. Error
Hospital	7774	3.21	0.014
No Hospital	90049	3.93	0.003

- 
- Q1: what is the expected health status of someone who has received hospitalization? (statistical prediction)
  - Q2: what will my health status be if I receive hospitalization? (causal prediction)

# Advertising and Sales



- Q1: what is the expected sales of a company with a given amount of TV ad spending? (statistical prediction)
- Q2: how much will my sales increase if I increase my TV ad spending by a certain amount? (causal prediction)

# Causal Effect Learning

In the simplest sense,

Causal effect learning = Statistical learning with  $E [y|do(x)]$  as the target function or  $p(y|do(x))$  as the target distribution.

# Causal Effect Learning

- To learn  $f(x) = E[y|\text{do}(x)]$ , perhaps the simplest way is to “just do it”.
- Let  $a$  be a possible value of  $x$ . Randomly select individual units, set their  $x = a$ , and observe the resulting  $y$ . In this way, we can generate data from  $p(y|\text{do}(x))$ .
  - ▶ This is in essence what a **randomized experiment** does.
- A nonparametric estimator for  $f(x)$  is then

$$\hat{f}(x = a) = \text{Ave}(y|x = a)$$

## Randomized Experiment

- Consider  $x \in \{0, 1\}$ . Suppose we are interested in learning the causal effect of  $x = 1$  on  $y$ .
- Given a set of experimental units, a **randomized controlled trial (RCT)** randomly selects a subset of individual units – call them the **treatment group** – to receive  $x = 1$ , and assign  $x = 0$  to the rest of the experimental units – called them the **control group**.

## Randomized Experiment

Using the experimental language,  $x$  is called **treatment** and  $y$  is called **outcome**. The **average treatment effect (ATE)** is defined as

$$\begin{aligned} \text{ATE} &= E[y|\text{do}(x=1)] - E[y|\text{do}(x=0)] \\ &\stackrel{[1]}{=} E[y|x=1] - E[y|x=0] \end{aligned}$$

, where [1] follows because randomized experiment generates data from  $p(y|\text{do}(x))$ , therefore  $E[y|x] = E[y|\text{do}(x)]$ .

For data generated by randomized experiments, correlation = causation.

# Randomized Experiment

## The Design of Experiments

By

Sir Ronald A. Fisher, Sc.D., F.R.S.

Honorary Research Fellow, Division of Mathematical Statistics, C.S.I.R.O., University of Adelaide; Foreign Associate, United States National Academy of Sciences; and Foreign Honorary Member, American Academy of Arts and Sciences; Foreign Member of the Swedish Royal Academy of Sciences, and the Royal Danish Academy of Sciences and Letters; Member of the Pontifical Academy; Member of the German Academy of Sciences (Leopoldina); formerly Galton Professor, University of London, and Arthur Balfour Professor of Genetics, University of Cambridge.



HAFNER PRESS  
A DIVISION OF MACMILLAN PUBLISHING CO., INC.  
New York  
COLLIER MACMILLAN PUBLISHERS  
London



SCIENCEPHOTOLIBRARY

## Randomized Experiment

- The Tennessee STAR project: randomly assigned students to classes of different sizes.
  - ▶ Naive contrasts suggest that children in larger classes have higher test scores.
  - ▶ Experiment result: children randomly assigned to smaller classes have higher test scores (Krueger 1999).
- Banerjee et al. (2010): randomly set up immunization camps in Indian villages and randomly offered rewards for receiving immunization.
  - ▶ Result: immunization camps with rewards substantially increased the rate of immunization among villagers.

# Randomized Experiment

- Two utility companies<sup>6</sup> randomly assigned customers to the treatment of receiving monthly or quarterly feedback on how much their energy consumption compared with their neighborhood average.
  - ▶ Result: households consuming more than neighborhood average tended to cut their consumption (Ayres, et al. 2009)<sup>7</sup>.
- List and Lucking-Reiley (2002): randomly assigned potential donors to fundraising campaigns with different amounts of seed money.
  - ▶ Result: more seed money sharply increased both the participation rate of donors and the average gift size received from participating donors.

---

<sup>6</sup>The Sacramento Municipal Utility District (SMUD) & Puget Sound Energy (PSE)

<sup>7</sup>Gerber et al. (2008) reports similar results for voter turnout.

# Observational Studies

- For many problems, RCTs are impossible or impractical to run:
  - ▶ ethical reasons
  - ▶ cost and duration
  - ▶ high-dimensionality
- How do we learn  $E[y|\text{do}(x)]$  from observational data?

# Observational Studies

- For observational data, correlation no longer implies causation. Consider the example of hospitalization and health: the fact that hospitalization is associated with worse health outcomes may not be due to the adverse effect of hospitalization on health, but to the fact the people with worse health received hospitalization in the first place. This is called **self-selection effect** or **self-selection bias**.
- Self-selection bias is a central concern to causal inference based on observed socio-economic data generated by individual choices.
- When individuals choose their own treatments (*self-selection*), those who choose to receive a treatment can be *systematically* different than those who choose not to, leading to a correlation between treatment and outcome that is not due to direct causation.

# Causal Model

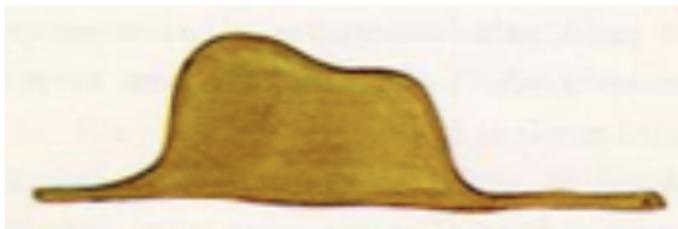
- To learn causal effects from observational data, we need an understanding of the **causal mechanism** that generates the data<sup>8</sup>.
- Causal mechanisms are described by causal models. A **causal model** for a set of random variables  $\{x_1, \dots, x_n\}$  is a model that specifies the joint distribution  $p(x_1, \dots, x_n)$ , as well as the causal relationships governing  $\{x_1, \dots, x_n\}$ .
- **Causal diagrams** are graphs that can be used to represent causal relationships and therefore describe our **qualitative** knowledge about a causal mechanism.

---

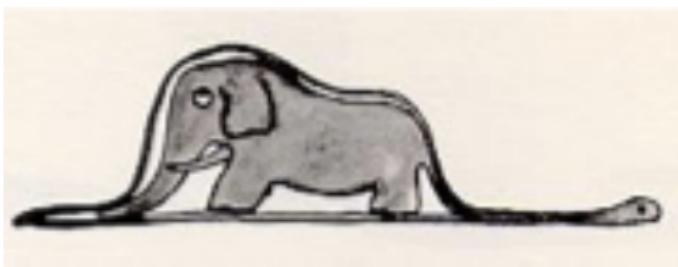
<sup>8</sup>In fact, as we will learn, such understanding is also necessary for interpreting and using experimental results. Without an understanding of – or making assumptions on – the underlying causal mechanism, any causal effect estimate is meaningless!

## Causal Diagrams

For example, suppose we see a boa constrictor that looks like this:



Our theory of the causal mechanism that leads to the boa constrictor looking like this is that it has just swallowed a baby elephant:



## Causal Diagrams

How do we represent our theory? We can write out a full causal model:

$$\log h^e \sim \mathcal{N}(0, 0.1), \log \ell^e \sim \mathcal{N}(0.5, 0.1), \log \ell^b \sim \mathcal{N}(1.5, 0.2)$$

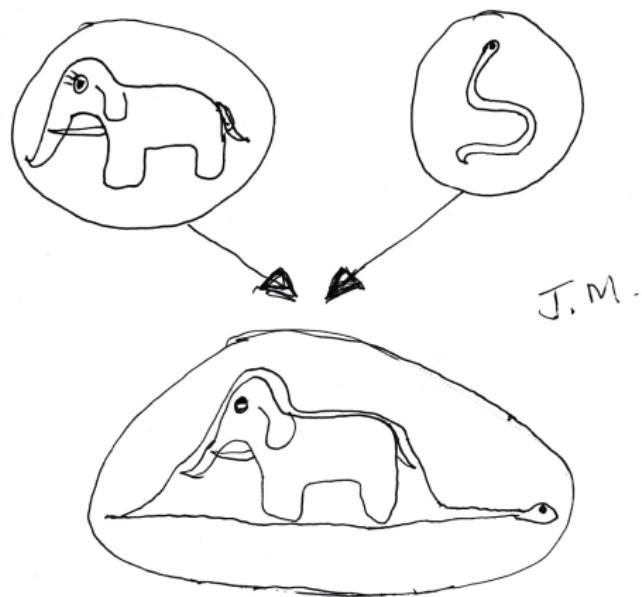
$$a | \ell^e, \ell^b, \ell^b > \ell^e \sim U(0, \ell^b - \ell^e)$$

$$y \stackrel{[1]}{=} \begin{cases} h^e \mathcal{I}(a \leq x \leq a + \ell^e) \mathcal{I}(E = 1) & \ell^e < \ell^b \\ 0 & \ell^e \geq \ell^b \end{cases}$$

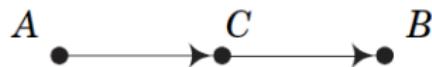
, where [1] is a **structural equation** with the *causal* meaning that the variable on the left is determined by the variables on the right,  $y$  is the height of the boa constrictor,  $x$  is the distance along the body of the boa constrictor from its head,  $(h^e, \ell^e)$  are respectively the height and length of the baby elephant,  $\ell^b$  is the length of the boa constrictor, and  $E \in \{0, 1\}$  is the event that the boa constrictor has swallowed the elephant.

# Causal Diagrams

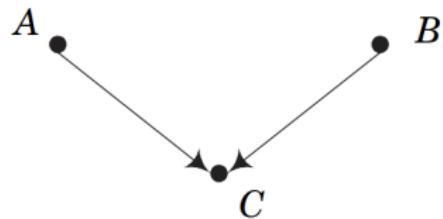
Or we can draw a causal diagram:



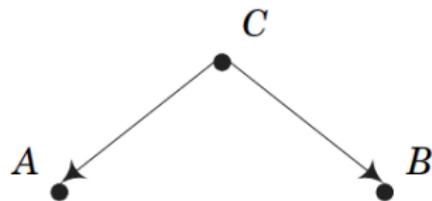
# Basic Patterns of Causal Relations



(a) Mediation



(c) Mutual causation



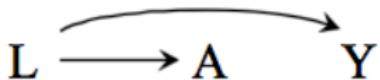
(b) Mutual dependence

Basic patterns of causal relationships among three variables

# Independence in Causal Diagrams

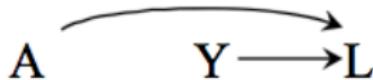
- A **collider** is a node  $c$  on a path with neighbors  $a$  and  $b$  such that  $a \rightarrow c \leftarrow b$ .
- A path is said to be **blocked** if it contains a noncollider that has been conditioned on, or it contains a collider that has not been conditioned on and has no descendants that have been conditioned on.
- Two variables are said to be **d-separated** if all paths between them are blocked. Otherwise they are **d-connected**.
- If two variables are d-separated (after conditioning on a set of variables), then they are (conditionally) independent.

## Association and Causation



- $L$  has a causal effect on both  $A$  and  $Y$ .  $A$  does not have a causal effect on  $Y$ .  $A$  depends on  $L$  and on *no other causes* of  $Y$ .
- $L$  is called a **common cause** to  $A$  and  $Y$ .
- $A$  and  $Y$  are **d-connected** and hence **associated**, because there exists an *open path*,  $A \leftarrow L \rightarrow Y$ , between them.
- Having information about  $A$  improves our ability to predict  $Y$ , even though  $A$  does not have a causal effect on  $Y$ .
- E.g.,  $A$  : carrying a lighter;  $Y$  : lung cancer;  $L$  : smoking

## Association and Causation



- Both  $A$  and  $Y$  have a causal effect on  $L$ .  $A$  does not have a causal effect on  $Y$ .
- $L$  is called a **common effect** of  $A$  and  $Y$ .
- $L$  is a collider on the path  $A \rightarrow L \leftarrow Y$  that has not been conditioned on. Hence  $L$  blocks the path.
- $A$  and  $Y$  are **d-separated** and hence **independent**, because the only path between them,  $A \rightarrow L \leftarrow Y$ , is blocked.
- E.g.,  $A$  : family heart disease history;  $Y$  : smoking;  $L$  : heart disease

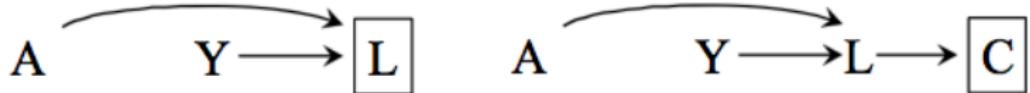
## Association and Causation



Box indicates conditioning

- Conditioning on  $B$  and  $L$  block the paths  $A \rightarrow B \rightarrow Y$  and  $A \leftarrow L \rightarrow Y$ .
- $A$  and  $Y$  are **d-separated** after conditioning on  $B$  and  $L$ . Therefore, they are **conditionally independent** given  $B$  and  $L$ , even though they are marginally associated in both graphs.
- E.g. (left),  $A$  : smoking;  $B$  : tar deposits in lung;  $Y$  : lung cancer

# Association and Causation



- Conditioning on collider  $L$  or its descendent  $C$  *opens* the path  $A \rightarrow L \leftarrow Y$ , which is blocked otherwise.
- $A$  and  $Y$  are **d-connected** after conditioning on  $L$  and  $C$ . Therefore, they are **conditionally associated** given  $L$  and  $C$ , even though they are marginally independent.
- E.g. (right),  $A$  : family heart disease history;  $Y$  : smoking;  $L$  : heart disease;  $C$  : taking heart disease medication

# Association and Causation

In summary, there are three structural reasons why two variables may be associated:

- ① One causes the other<sup>9</sup>
- ② They share common causes
- ③ The analysis is conditioned on their common effects<sup>10</sup>

---

<sup>9</sup>either directly or through mediating variables, i.e. there exists an open causal path that connects the two variables.

<sup>10</sup>or the consequences of the common effects.

# Confounding

- When two variables share common causes, they are correlated even if they do not cause each other. This makes it harder for us to learn the causal effect one has on the other. We call this problem **confounding**. The common causes are called **confounders**.
- Self-selection bias is an important type of confounding. Consider treatment  $x$  and outcome  $y$ . When  $x$  is selected based on the values of  $z$ , if  $z$  also has a causal effect on  $y$ , then  $z$  is a confounder and there is self-selection bias.

# The Back-Door Strategy

- Therefore, to learn the causal effect of  $x$  on  $y$ , we can condition on their common causes<sup>11</sup>, while avoiding to condition on any of their common effects.
  - ▶ Conditioning on common causes make two variables independent if they do not have causal effects on each other.
  - ▶ Therefore, any association between two variables after their common causes have been conditioned on should be due to causation.
  - ▶ This is in essence the “back-door strategy”.

---

<sup>11</sup>When we condition on a variable, we also say we **control for** the variable.

# The Back-Door Strategy

- A **back-door path** is a path between  $x$  and  $y$  that has an arrow *into*  $x$ : These are the paths that, if left open, induce association between  $x$  and  $y$  that is not a result of  $x$  causing  $y$ .
- To learn the causal effect of  $x$  on  $y$ , the **back-door strategy** is to block all back-door paths from  $x$  to  $y$ . Formally, let  $z$  be a set of variables such that (i) conditioning on  $z$  blocks every back-door path from  $x$  to  $y$ , and (ii)  $z$  does not contain any descendant of  $x$ . Then

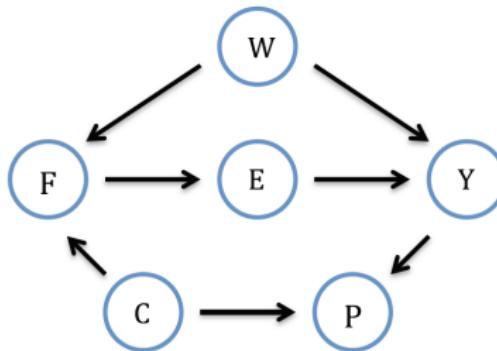
$$p(y|\text{do}(x), z) = p(y|x, z)$$

$$p(y|\text{do}(x)) = \sum_z p(y|x, z) p(z)$$

## Fumigation and Yield

- Fumigation is the use of fumigants to control eelworms. We want to know the effect of fumigation on crop yield.
- Suppose the mechanism that generates our observed data works as follows: fumigation ( $F$ ) helps control eelworms ( $E$ ), which affects crop yield ( $Y$ ). Farmers' use of fumigation is affected by weather ( $W$ ) and the price of fumigants ( $C$ ). Weather also affects yield independently. Finally, the equilibrium crop price ( $P$ ) is affected by both the price of fumigants and the realized crop yield.

## Fumigation and Yield



---

The back-door path  $F \leftarrow C \rightarrow P \leftarrow Y$  is already blocked by the collider  $P$ . Therefore, using the back-door strategy, we only need to control for  $W$ :

$$p(Y|\text{do}(F)) = \sum_W p(Y|F, W) p(W)$$

# Smoking and Child Birthweight

- Want to know: the effect of mother's smoking behavior on child birthweight.
- Problem: age ( $A$ ) is a potential confounder, since mothers' smoking behavior ( $S$ ) varies by age and age directly affects child birthweight ( $W$ ).
- Solution: using the back-door strategy<sup>12</sup>,

$$E[W|\text{do}(S), A] = E[W|S, A]$$

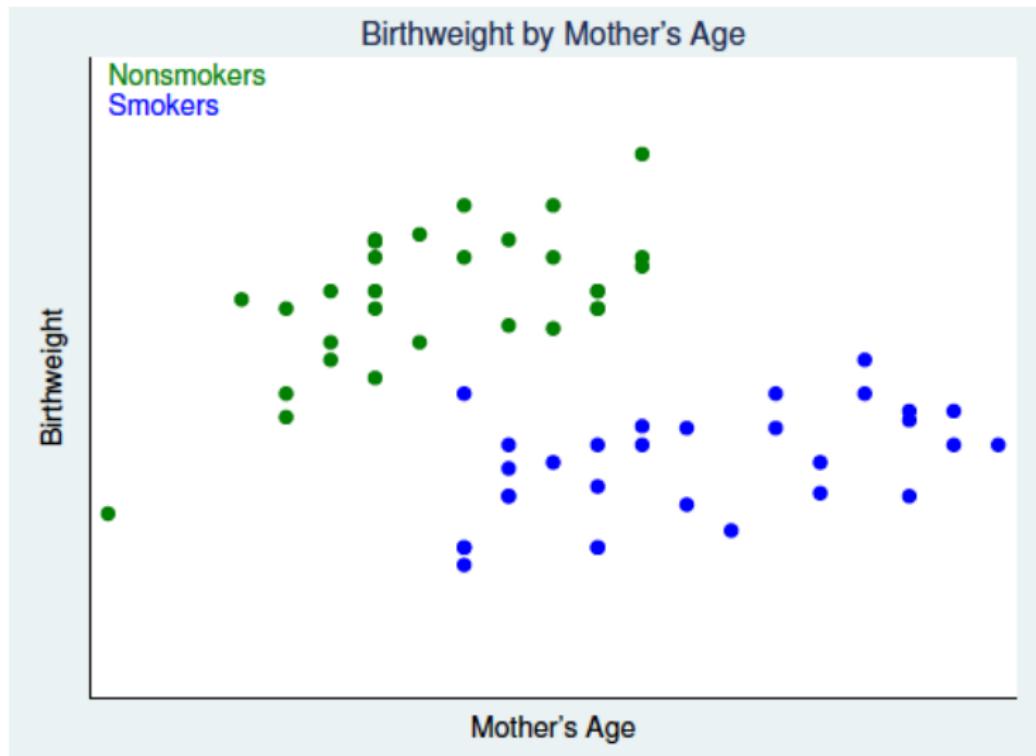
$$E[W|\text{do}(S)] = \sum_A E[W|S, A] p(A)$$

---

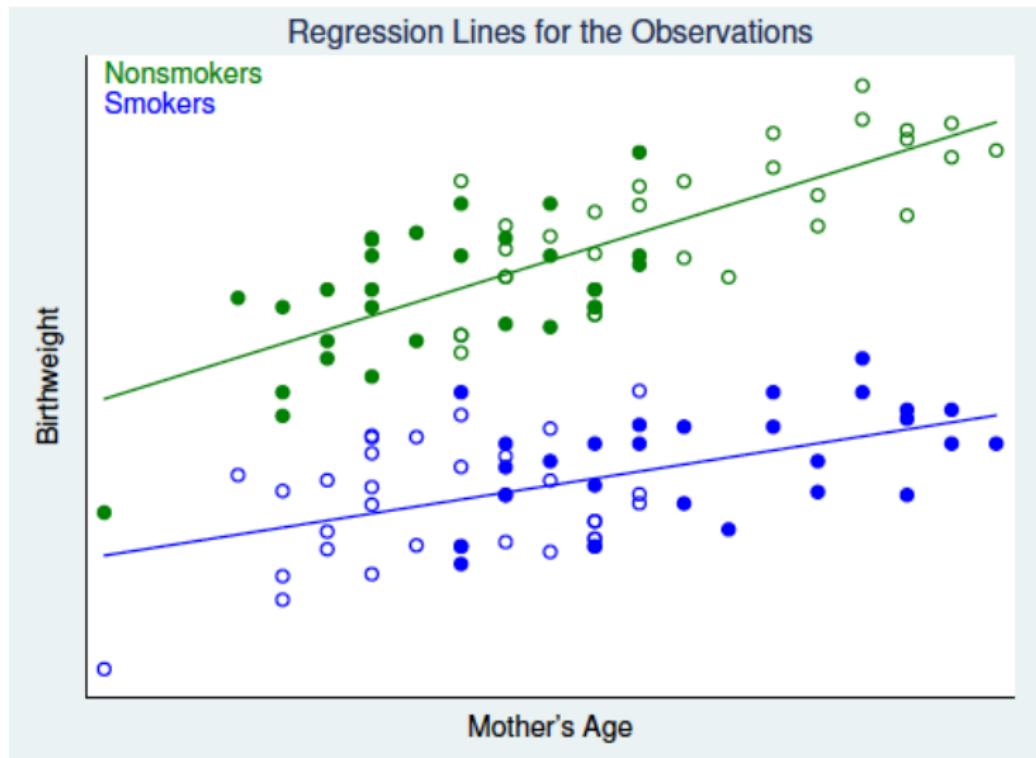
<sup>12</sup>We can, for example, fit a linear model to  $E[W|S, A]$ :

$$E[W|\text{do}(S), A] = E[W|S, A] = \beta_0^S + \beta_1^S A$$

# Smoking and Child Birthweight



# Smoking and Child Birthweight



## Structual Estimation

In the econometrics literature, causal models based on **economic theory** are referred to as **structural models**. Their estimation is called **structural estimation**.

- Structural models use economic theory to specify the *functional form* of causal relationships.
- They are a type of *scientific models*.

# Structual Estimation

## Scientific vs. Statistical Model

If you want to predict where Mars will be in the night sky<sup>a</sup>, you may do very well with a model in which Mars revolves around the Earth. You can estimate, from data, how fast Mars goes around the Earth and where it should be tonight. But the estimated model does not describe the actual causal mechanisms.

---

<sup>a</sup>This example is taken from Shalizi (2016).

# Auction



First-price Sealed-bid Auctions for Identical Goods

# Auction

## Model

- $N$  risk-neutral bidders
- Independent private value  $v_i \sim i.i.d. F(.)$
- Each bidder knows her own  $v_i$  and the distribution  $F$ , but not the  $v_i$  of others
- Observed bids are the Bayesian Nash equilibrium outcome of the game

⇒ Equilibrium bidding strategy:

$$\begin{aligned} b_i &= v_i - \frac{1}{F(v_i)^{N-1}} \int_0^{v_i} F(x)^{N-1} dx \\ &= v_i - \frac{1}{N-1} \frac{G_N(b_i)}{g_N(b_i)} \end{aligned}$$

, where  $G_N(.)$  and  $g_N(.)$  are the c.d.f. and p.d.f. of the bid distribution.

# Auction

## Structural Estimation

- ① For each auction<sup>a</sup>, nonparametrically estimate  $G_N(\cdot)$  and  $g_N(\cdot)$  from observed bids  $\{b_1, \dots, b_N\}$ .
- ② For each bidder, calculate

$$\hat{v}_i = b_i + \frac{1}{N-1} \frac{\hat{G}_N(b_i)}{\hat{g}_N(b_i)} \quad (1)$$

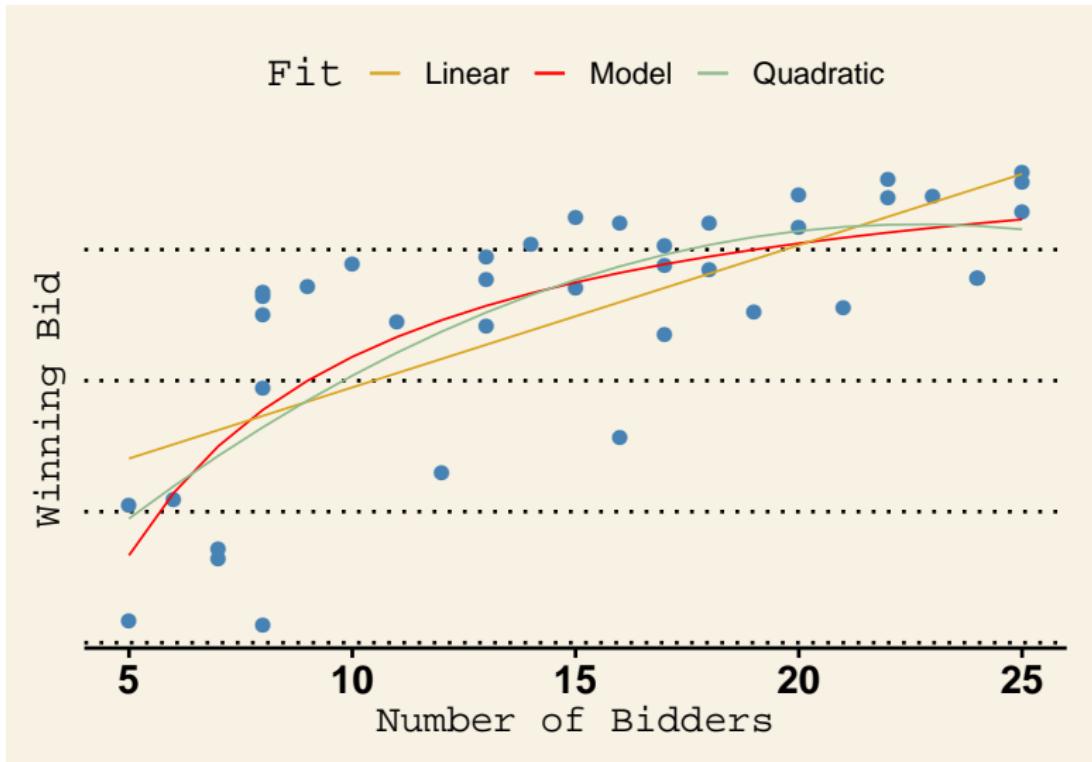
- ③ Use  $\hat{v}_i$  to nonparametrically estimate  $F(\cdot)$
- ④  $\hat{F}(\cdot)$  can be used to predict the winning bid in an  $N$ -bidder auction:

$$E[\max\{b_i\}] = E \left[ \max \left\{ v_i - \frac{1}{\hat{F}(v_i)^{N-1}} \int_0^{v_i} \hat{F}(x)^{N-1} dx \right\} \right]$$

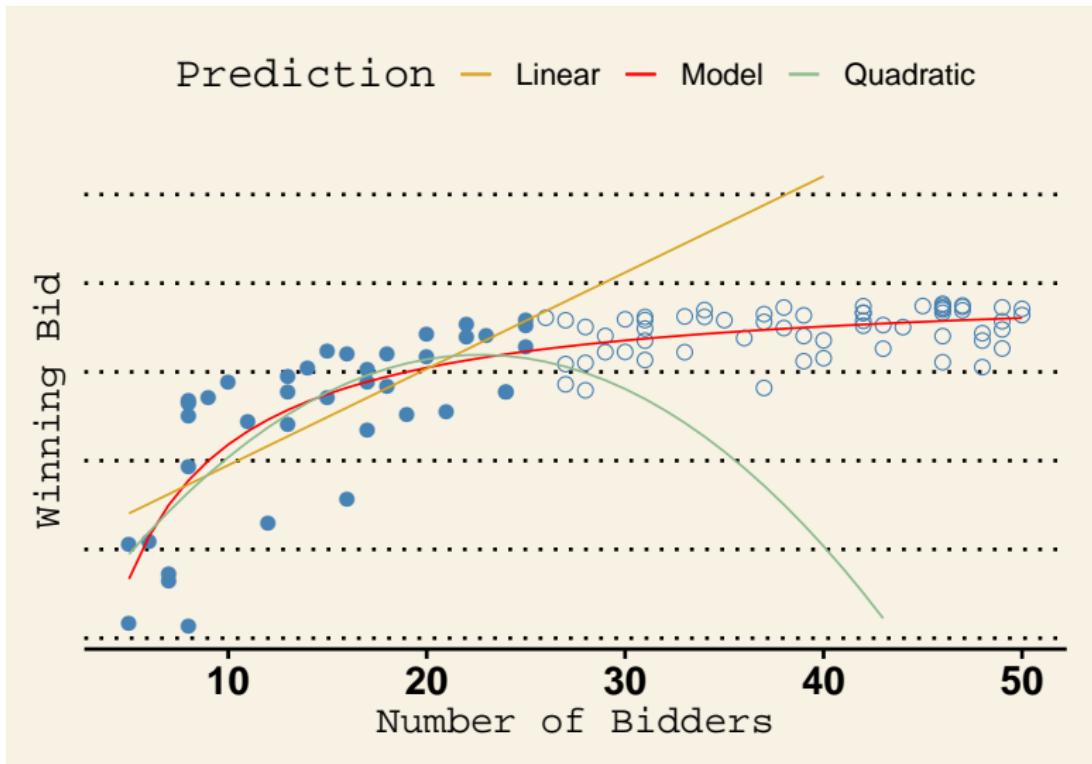
---

<sup>a</sup>See Guerre et al. (2000).

# Auction



# Auction



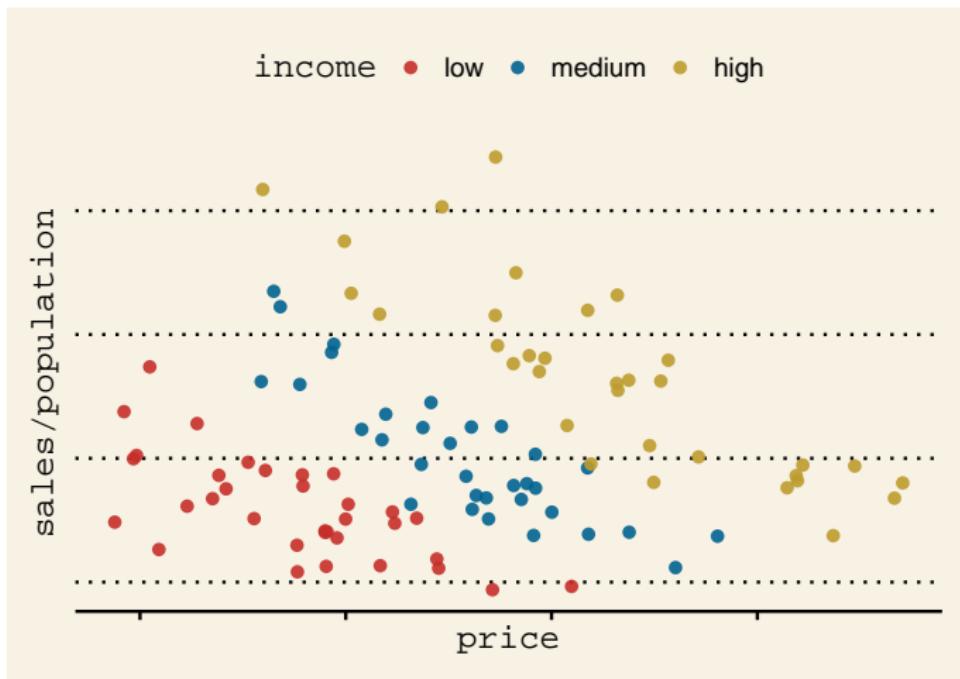
## Auction

- Here, there is no confounding between  $N$  (the number of bidders) and  $b_{\max}$  (the winning bid). Hence
$$f(N) \equiv E[b_{\max} | \text{do}(N)] = E[b_{\max} | N].$$
- The learning problem is to learn  $f(N)$  from data. Here, theory helps specify the functional form of  $f(N)$  and therefore serves as a **model selection** mechanism.
- Theory also helps us to learn the values of the bidders (equation (1)) by specifying the functional form of the mapping from  $\{v_i\}$  to  $\{b_i\}$ .

# Monopoly

A monopoly firm's pricing and sales in different geographical markets

Data: price, sales, average income, population for each market



# Monopoly

## Model: Demand

In each market  $m$  with population  $N_m$  and mean income  $I_m$ , consumers choose between the monopoly product and an outside good. Individual utilities are given by:

$$U_{i0}^m = \epsilon_{i0}^m \quad (2)$$

$$U_{i1}^m = \beta_0 + \beta_1 I_m - \beta_2 p_m + \epsilon_{i1}^m$$

, where  $(U_{i0}^m, U_{i1}^m)$  are respectively the indirect utilities of the outside good and the monopoly product, and  $\epsilon_{ij}^m \sim \text{Gumbel}(0, 1)$ .

(2)  $\Rightarrow q_m \sim \text{Binomial}(N_m, \pi_m)$ , where

$$\pi_m = \frac{\exp(\beta_0 + \beta_1 I_m - \beta_2 p_m)}{1 + \exp(\beta_0 + \beta_1 I_m - \beta_2 p_m)}$$

# Monopoly

## Model: Supply

For each market  $m$ , given demand  $q_m(p)$ , the monopoly firm chooses  $p$  to maximize:

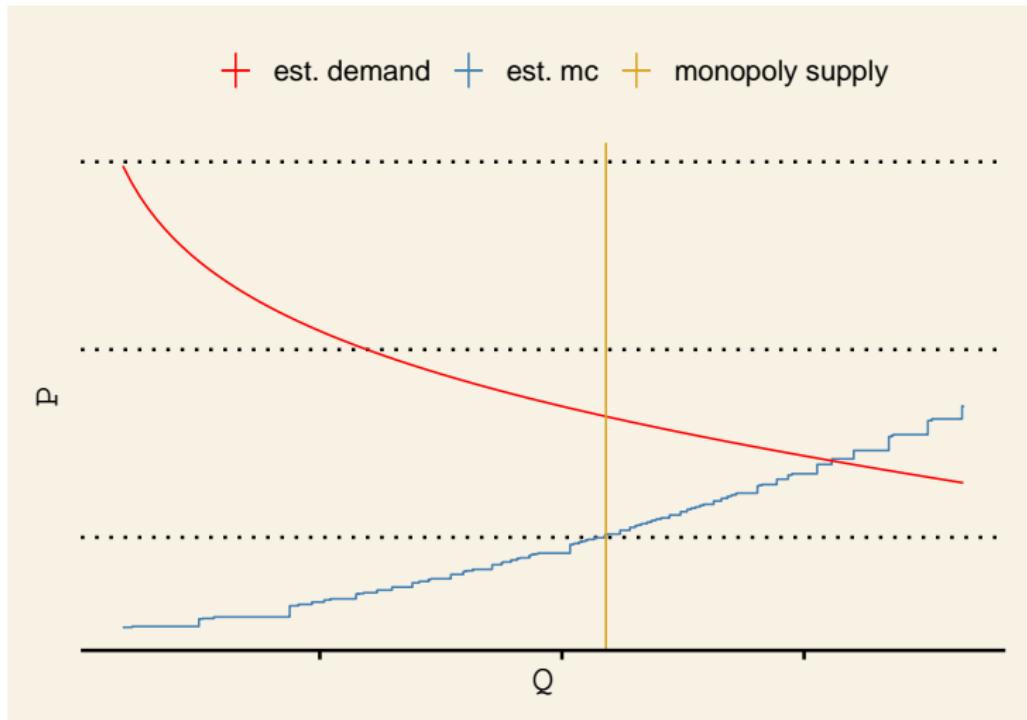
$$\max_p \{p \times q_m(p) - c(q_m(p))\} \quad (3)$$

, where  $c(q)$  is the firm's cost function.

(3)  $\Rightarrow$

$$c'(q_m) = p_m + [q'_m(p_m)]^{-1} q_m \quad (4)$$

# Monopoly



Estimated marginal cost and demand curves  
for a market with median income and population

# Monopoly

- Here, theory helps us to learn the marginal cost function of the monopoly firm as well as the consumer utility function.
- Using the estimation results, we can conduct **welfare analysis** and make **normative statements**.
  - ▶ For example, calculating the total deadweight loss due to monopoly.

## Counterfactual Simulation

- One of the benefits of learning a structural model is that it allows us to predict the effect of a completely new treatment – a treatment that has never been observed before.
  - ▶ If in the observed data,  $x_j$  is always equal to 0, what would be the effect of  $\text{do}(x_j = 1)$ ?
- Because structural estimation learns an entire structural model, once we have learned a model with variables  $\{x_1, \dots, x_n\}$ , we can use it to generate data  $\mathcal{D} = \{(x_{i,1}, \dots, x_{i,j-1}, x_{i,j} = a, x_{i,j+1}, x_{i,n})\}$  from the *post-intervention* distribution  $p(x_1, \dots, x_n | \text{do}(x_j = a))$  for any hypothetical manipulation  $\text{do}(x_j = a)$ . This is called **counterfactual simulation**.

# Counterfactual Simulation

## The Road Not Taken By Robert Frost

TWO roads diverged in a yellow wood,  
And sorry I could not travel both  
And be one traveler, long I stood  
And looked down one as far as I could  
To where it bent in the undergrowth;

Then took the other, as just as fair,  
And having perhaps the better claim,  
Because it was grassy and wanted wear;  
Though as for that the passing there  
Had worn them really about the same,

And both that morning equally lay  
In leaves no step had trodden black.  
Oh, I kept the first for another day!  
Yet knowing how way leads on to way,  
I doubted if I should ever come back.

I shall be telling this with a sigh  
Somewhere ages and ages hence:  
Two roads diverged in a wood, and I—  
I took the one less traveled by,  
And that has made all the difference.



# Counterfactual Simulation



What if Chuji Qu never visited Liu's Village?

# Counterfactual Simulation



Or Caesar never crossed the Rubicon?

# Monopoly

What happens if the government imposes a 20% sales tax on the company?

After tax:

Δ Consumer Surplus: -27.83%  
Δ Total Surplus: -27.95%

Tax incidence:

Consumer: 26.65%

# Structual Estimation

- Successful structural models have the potential to deliver better predictive performance than statistical models trained on single data sets<sup>13</sup>, because their parameters can be learned from a combination of data from various sources that share the same underlying causal mechanisms.
  - ▶ Apples falling down trees and the earth orbiting around the sun both inform us of the gravitational constant.
  - ▶ Individuals' investment behavior and career choices can both inform us of their degrees of risk aversion.

---

<sup>13</sup>Think of quantum mechanics!

# Acknowledgement I

Part of this lecture is adapted from the following sources:

- Antoine de Saint-Exupéry. 2001. *Le Petit Prince*. Harcourt, Inc. (Original work published 1943.)
- Doré, G. *The Dore Illustrations for Dante's Divine Comedy*. Dover Publications, 1st edition (1976).
- Hernán, M. A. and J. M. Robins. 2018. *Causal Inference*. Boca Raton: Chapman & Hall/CRC, forthcoming.
- James, G., D. Witten, T. Hastie, and R. Tibshirani. 2013. *An Introduction to Statistical Learning: with Applications in R*. Springer.
- Morgan, S. L. and C. Winship. 2007. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge University Press.

## Acknowledgement II

- Pearl, J., M. Glymour, and N. P. Jewell. 2016. *Causal Inference in Statistics: A Primer*. Wiley.
- Silva, R. *Causal Inference in Machine Learning*. Talk at Imperial College London, retrieved on 2017.01.01. [[link](#)]
- StataCorp. 2013. *Stata Treatment-effects Reference Manual Release 13*. Stata Press.
- Varian, H. R., *Machine Learning and Econometrics*. Talk at Google, retrieved on 2017.01.01. [[link](#)]

# Reference I

-  Ayres, I., S. Raseman, and A. Shih. 2013. "Evidence from Two Large Field Experiments that Peer Comparison Feedback Can Reduce Residential Energy Usage," *Journal of Law, Economics, & Organization*, 29(5).
-  Banerjee, A. V., E. Duflo, R. Glennerster, and D. Kothari. 2010. "Improving immunisation coverage in rural India: clustered randomised controlled evaluation of immunisation campaigns with and without incentives," *BMJ*.
-  Gerber, A. S., D .P. Green, and C. W. Larimer. 2008. "Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment," *American Political Science Review*, 102(1).
-  Krueger, A. B. 1999. "Experimental Estimates of Education Production Functions," *Quarterly Journal of Economics*, 114(2).
-  List, J.A. and D. Lucking-Reiley. 2002. "Effects of seed money and refunds on charitable giving: experimental evidence from a university capital campaign," *Journal of Political Economy*, 110(1).

## Reference II



- Shalizi, C. R. 2016. *Advanced Data Analysis from an Elementary Point of View*. Manuscript.