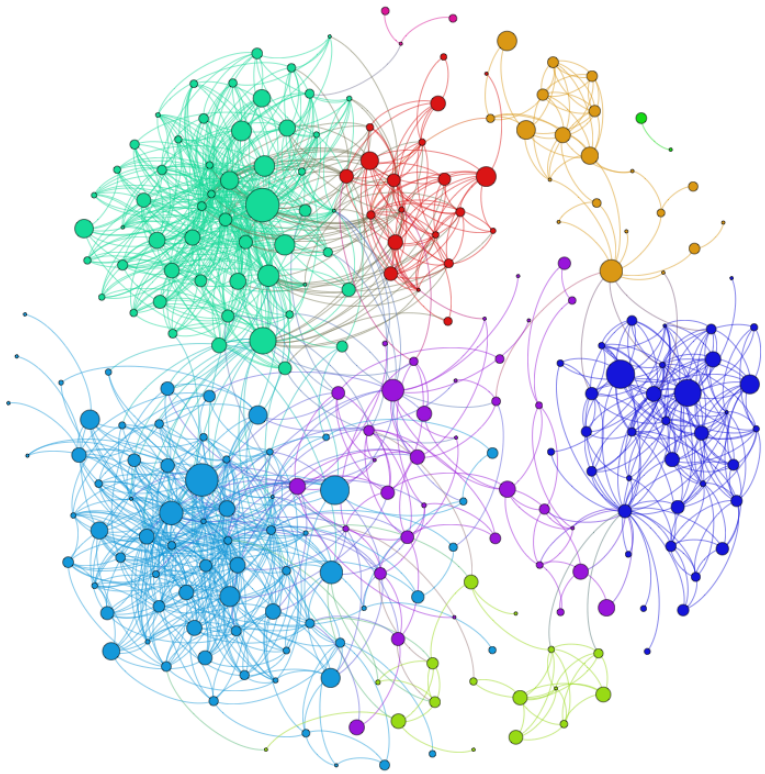
Data Analysis for Economics

ä Modern Introduction

Jiaming Mao

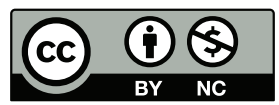
Xiamen University



This version: Spring 2019

Contact: jmao@xmu.edu.cn

Course homepage: jiamingmao.github.io/data-analysis



Data are everywhere

	Cheese				
0.5/0.51 lb	**Cabot Vermont Cheddar**	0.51 lb	$7.99/lb	**$4.07**	
	Dairy				
1/1	**Friendship Lowfat Cottage Cheese** (16oz)		$2.89/ea	**$2.89**	
1/1	**Nature's Yoke Grade A Jumbo Brown Eggs** (1 dozen)		$1.49/ea	**$1.49**	
1/1	**Santa Barbara Hot Salsa, Fresh** (16oz)		$2.69/ea	**$2.69**	
1/1	**Stonyfield Farm Organic Lowfat Plain Yogurt** (32oz)		$3.59/ea	**$3.59**	
	Fruit				
3/3	**Anjou Pears** (Farm Fresh, Med)	1.76 lb	$2.49/lb	**$4.38**	
2/2	**Cantaloupe** (Farm Fresh, Med)		$2.00/ea	**$4.00**	**S**
	Grocery				
1/1	**Fantastic World Foods Organic Whole Wheat Couscous** (12oz)		$1.99/ea	**$1.99**	
1/1	**Garden of Eatin' Blue Corn Chips** (9oz)		$2.49/ea	**$2.49**	
1/1	**Goya Low Sodium Chickpeas** (15.5oz)		$0.89/ea	**$0.89**	
2/2	**Marcal 2-Ply Paper Towels, 90ct** (1ea)		$1.09/ea	**$2.18**	**T**
1/1	**Muir Glen Organic Tomato Paste** (6oz)		$0.99/ea	**$0.99**	
1/1	**Starkist Solid White Albacore Tuna in Spring Water** (6oz)		$1.89/ea	**$1.89**	

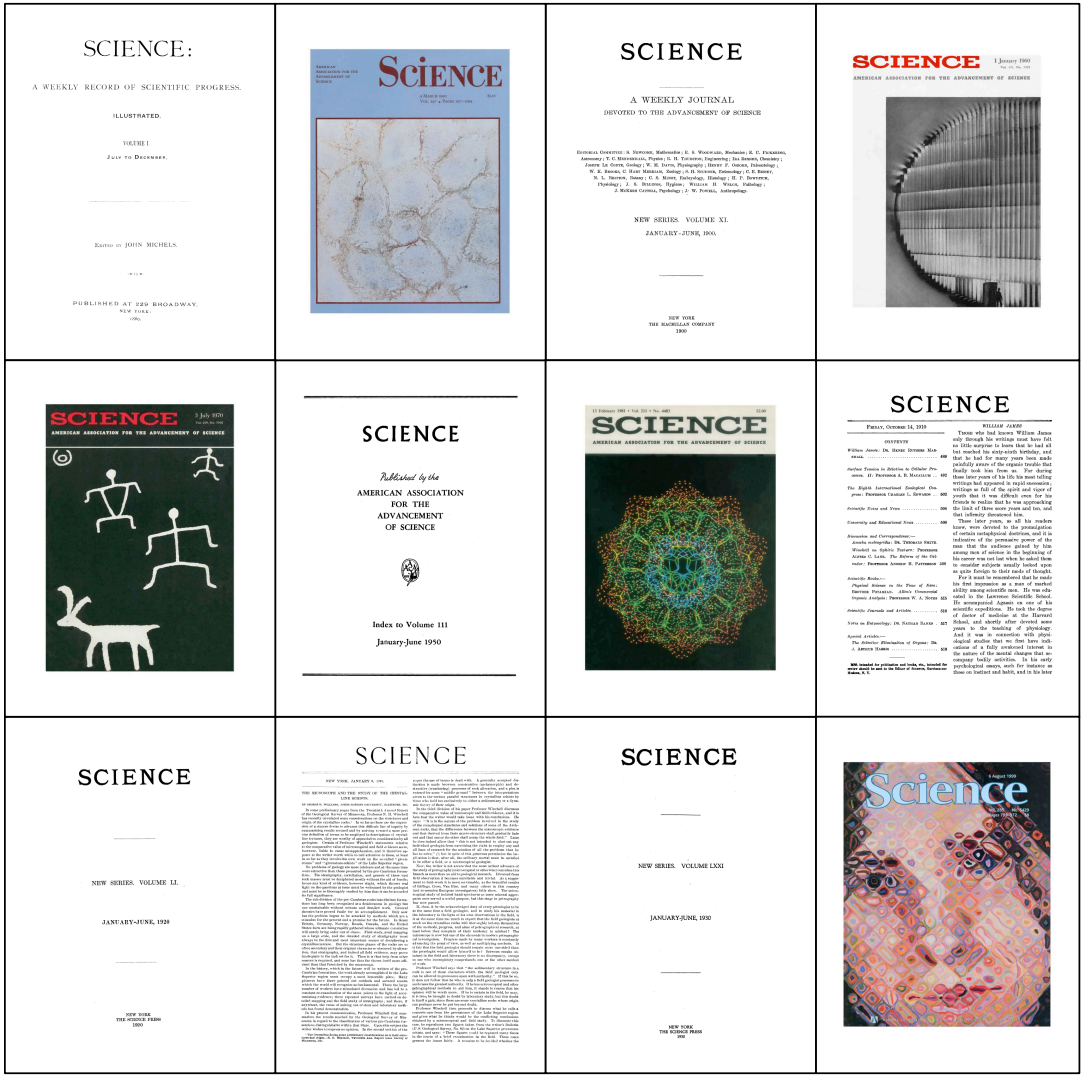
Purchase histories

Data are everywhere

Ikiru (1952)	UR	Foreign	
Junebug (2005)	R	Independent	
La Cage aux Folles (1979)	R	Comedy	
The Life Aquatic with Steve Zissou (2004)	R	Comedy	
Lock, Stock and Two Smoking Barrels (1998)	R	Action & Adventure	
Lost in Translation (2003)	R	Drama	
Love and Death (1975)	PG	Comedy	
The Manchurian Candidate (1962)	PG-13	Classics	
Memento (2000)	R	Thrillers	
Midnight Cowboy (1969)	R	Classics	

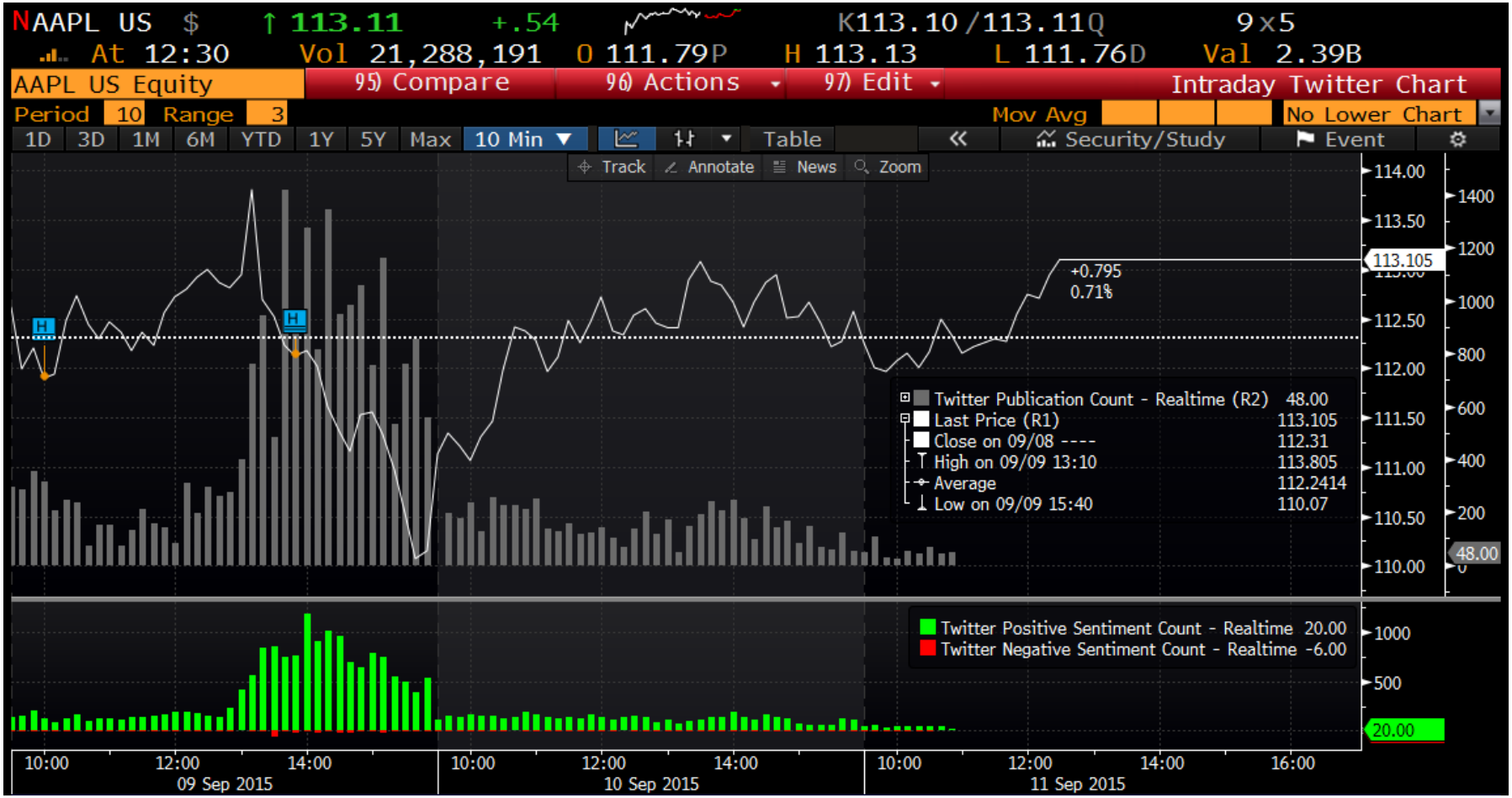
User ratings

Data are everywhere



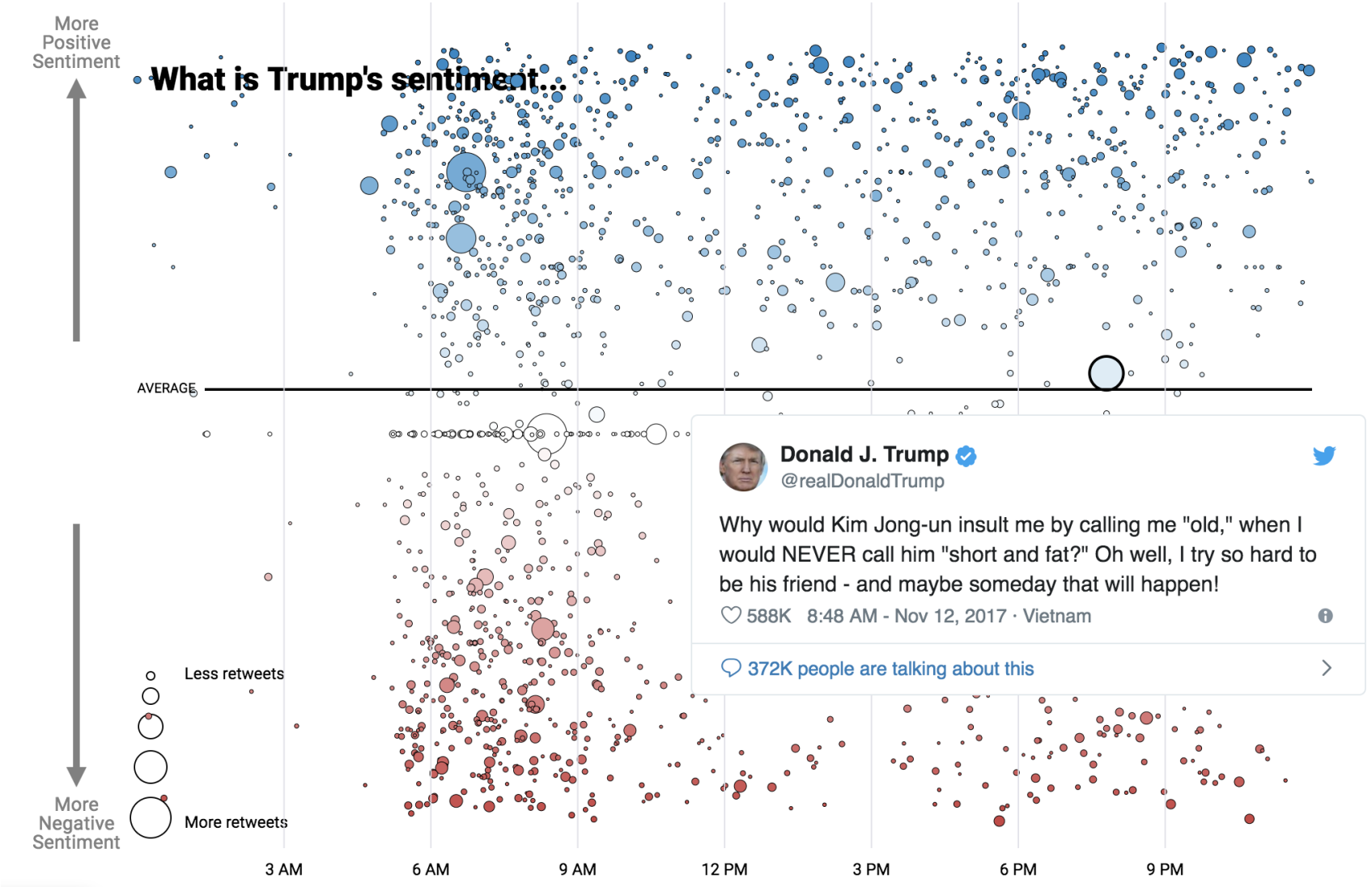
Document collections

Data are everywhere



Financial markets

Data are everywhere



Social networks

Data Science

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

Machine Learning $\rightarrow$ Statistics $\rightarrow$ 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.

Classification



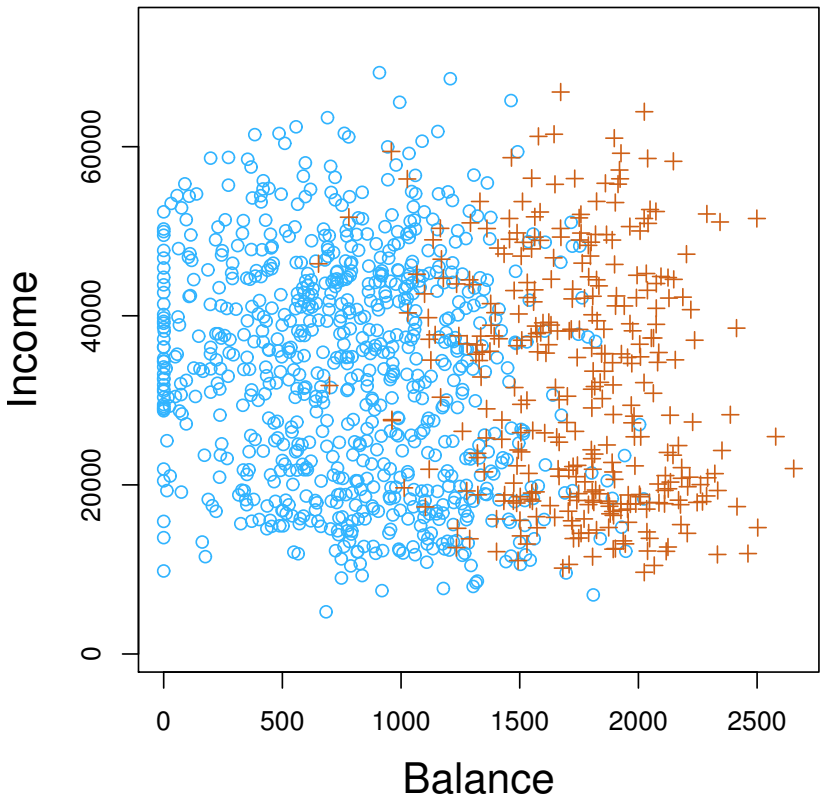


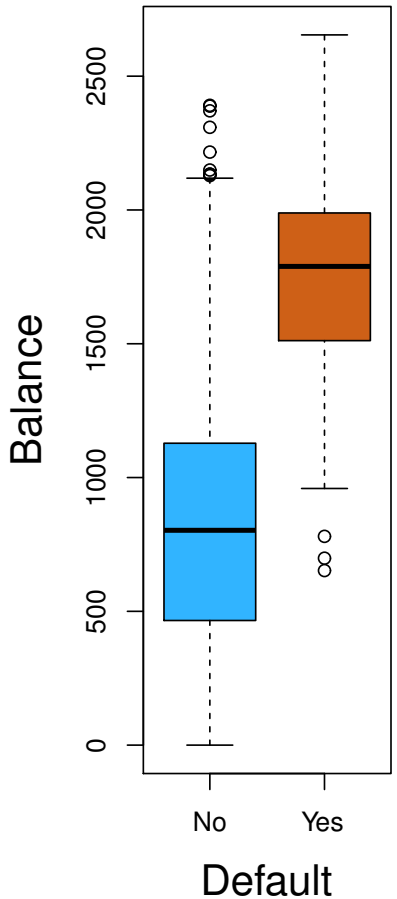


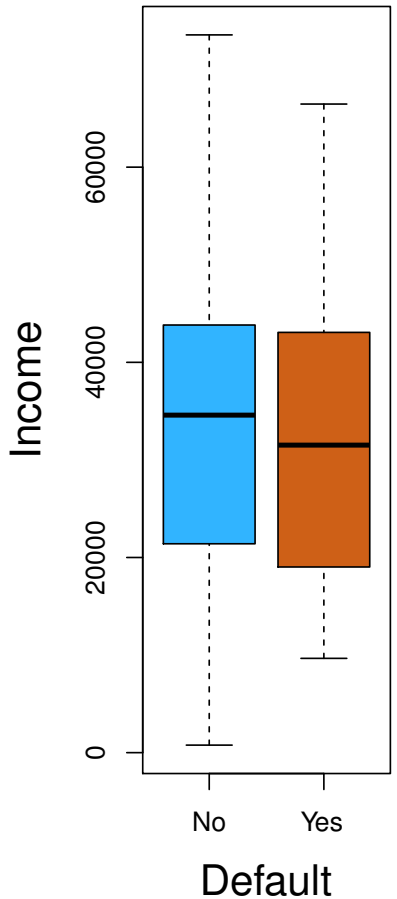


Which one is a chair?

Classification

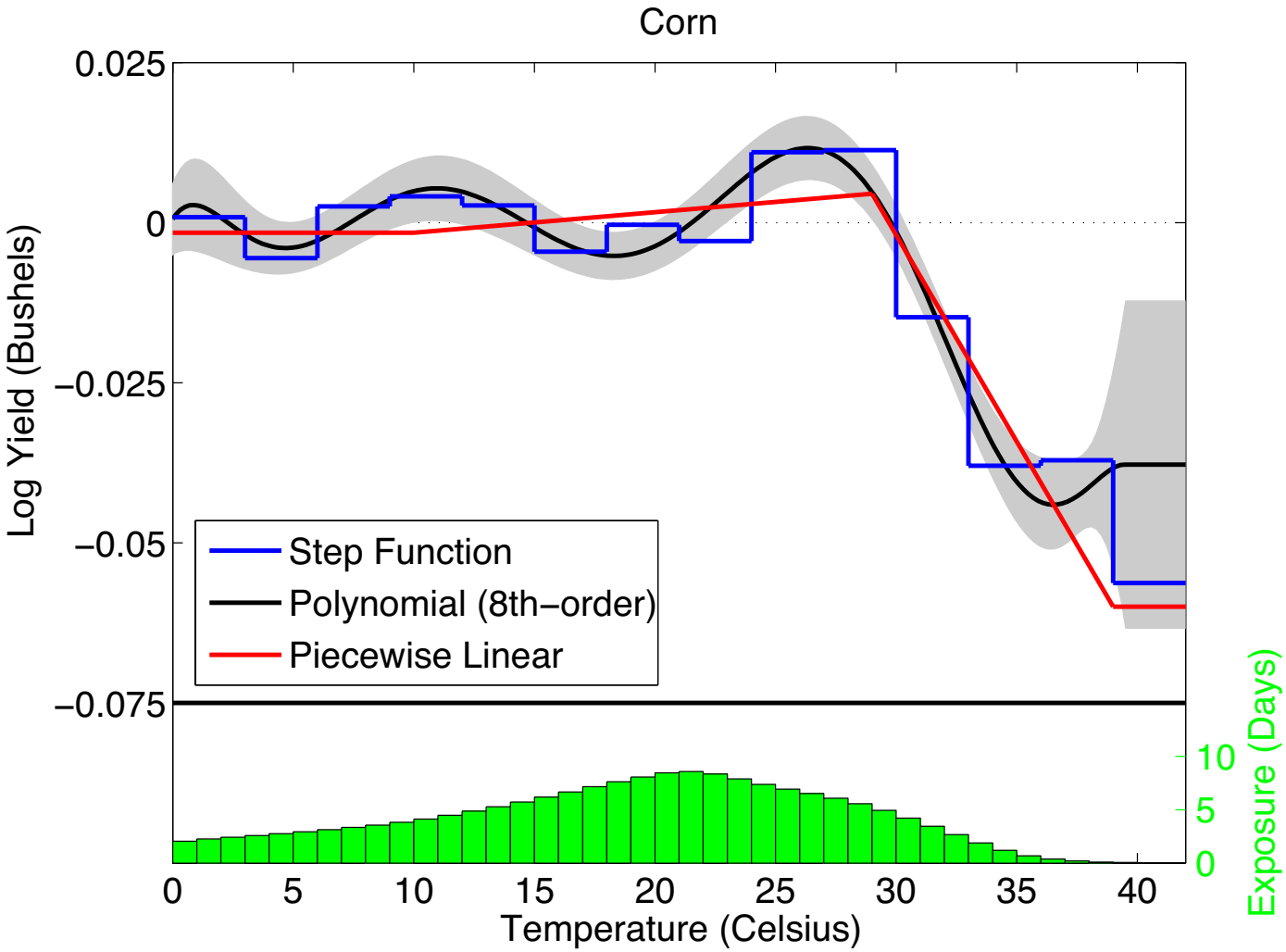






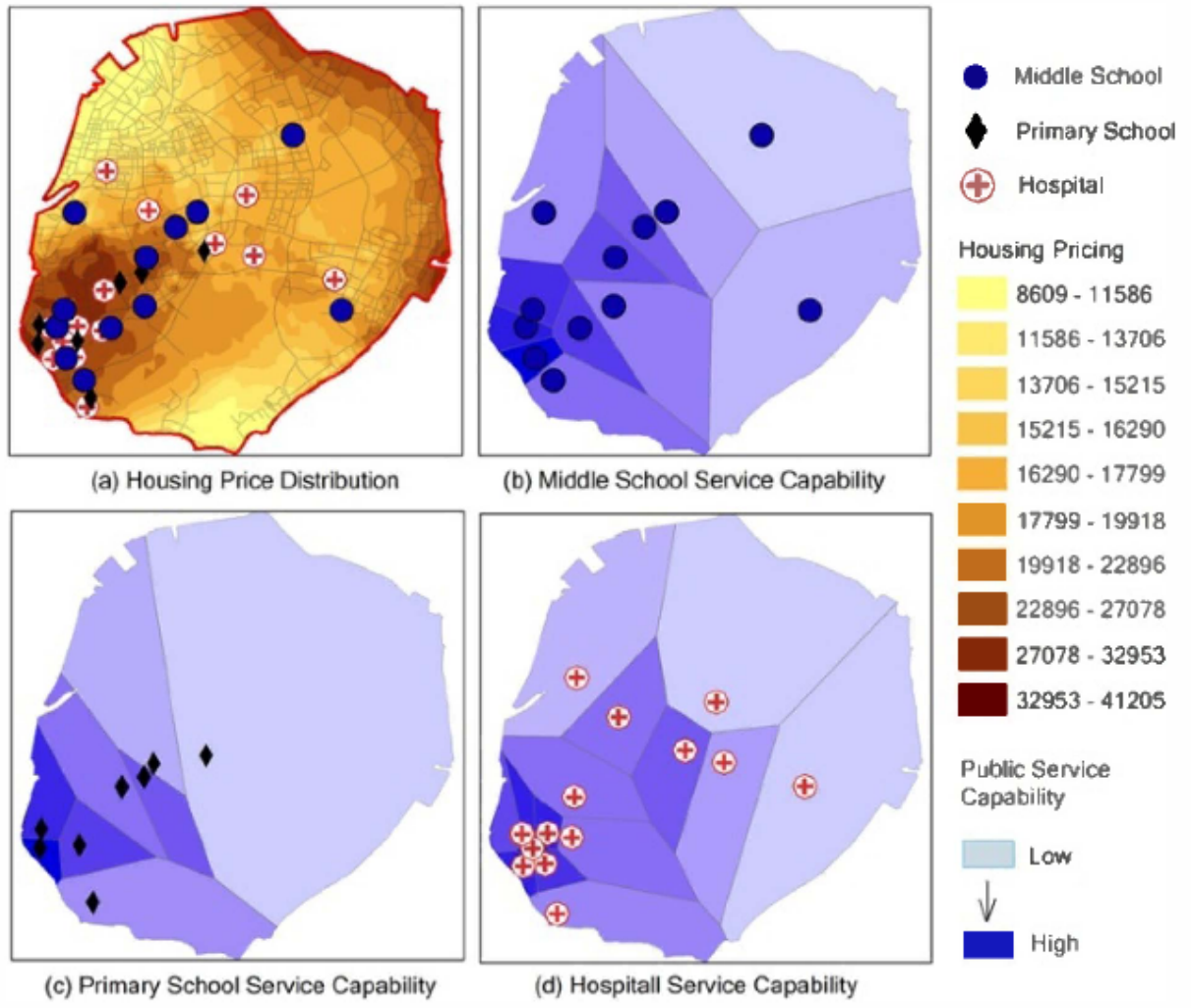
Who will default on their credit card?

Regression



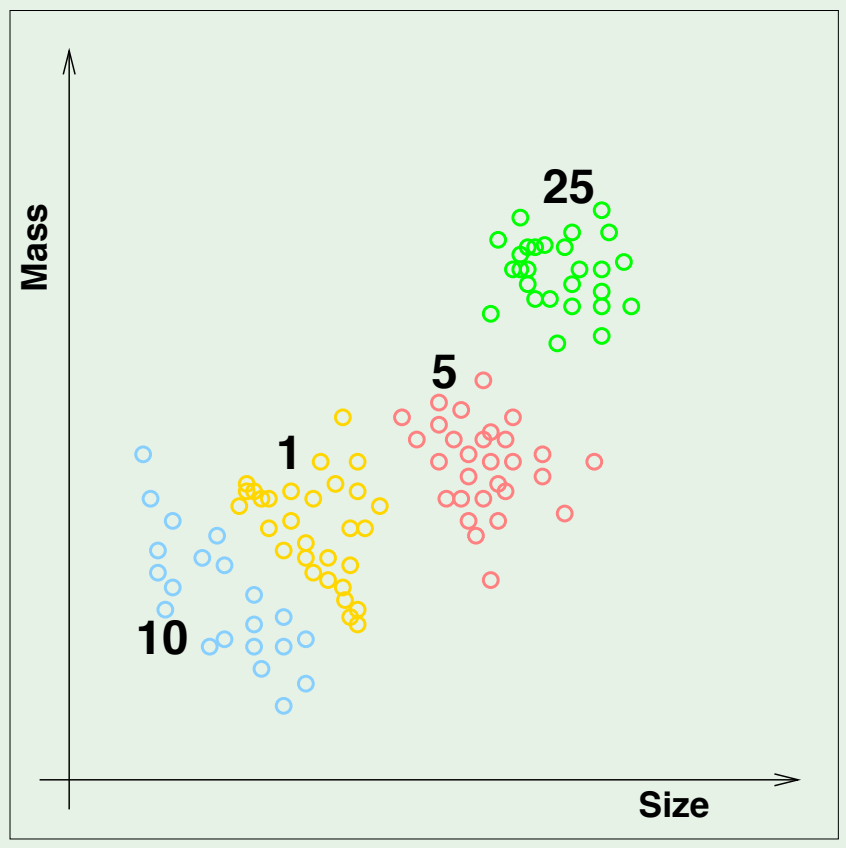
Temperature and crop yield

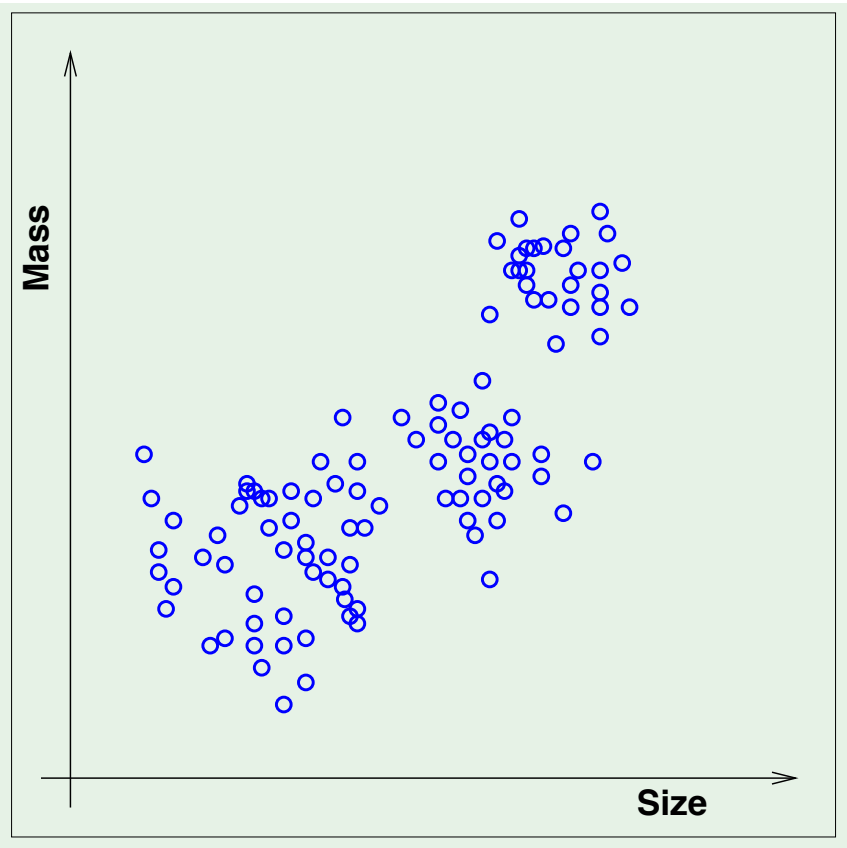
Regression



Amenities and housing prices (Xiamen, China)

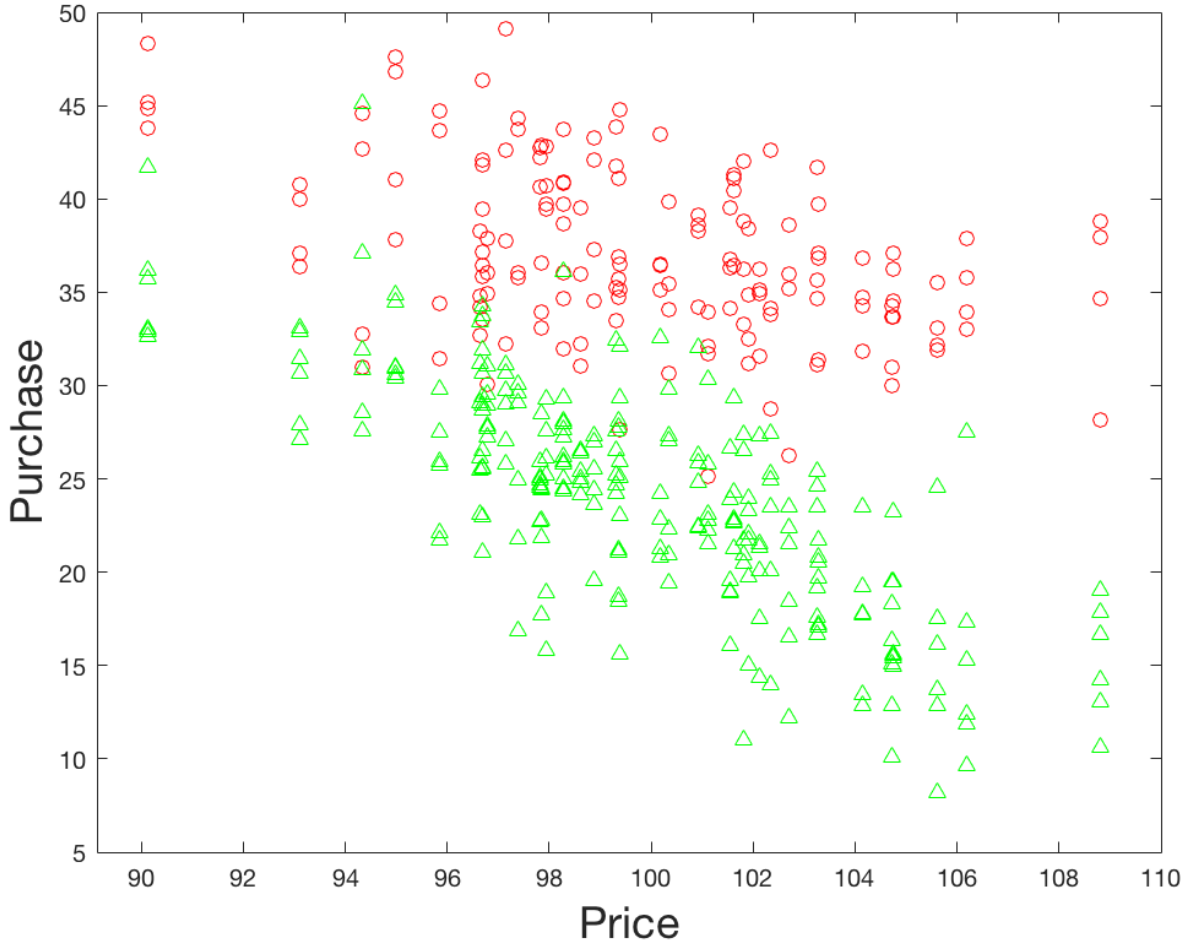
Supervised vs. Unsupervised Learning





Vending machine coin recognition
Left: supervised learning; Right: unsupervised learning

Supervised vs. Unsupervised Learning



Consumer demand

Causal Inference

Learning patterns in the data is not enough. We want understanding.

- The focus of ☼science☼.

Bertrand Russell[a] told the following cautionary tale of the perils of learning patterns without understanding:

> *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.*

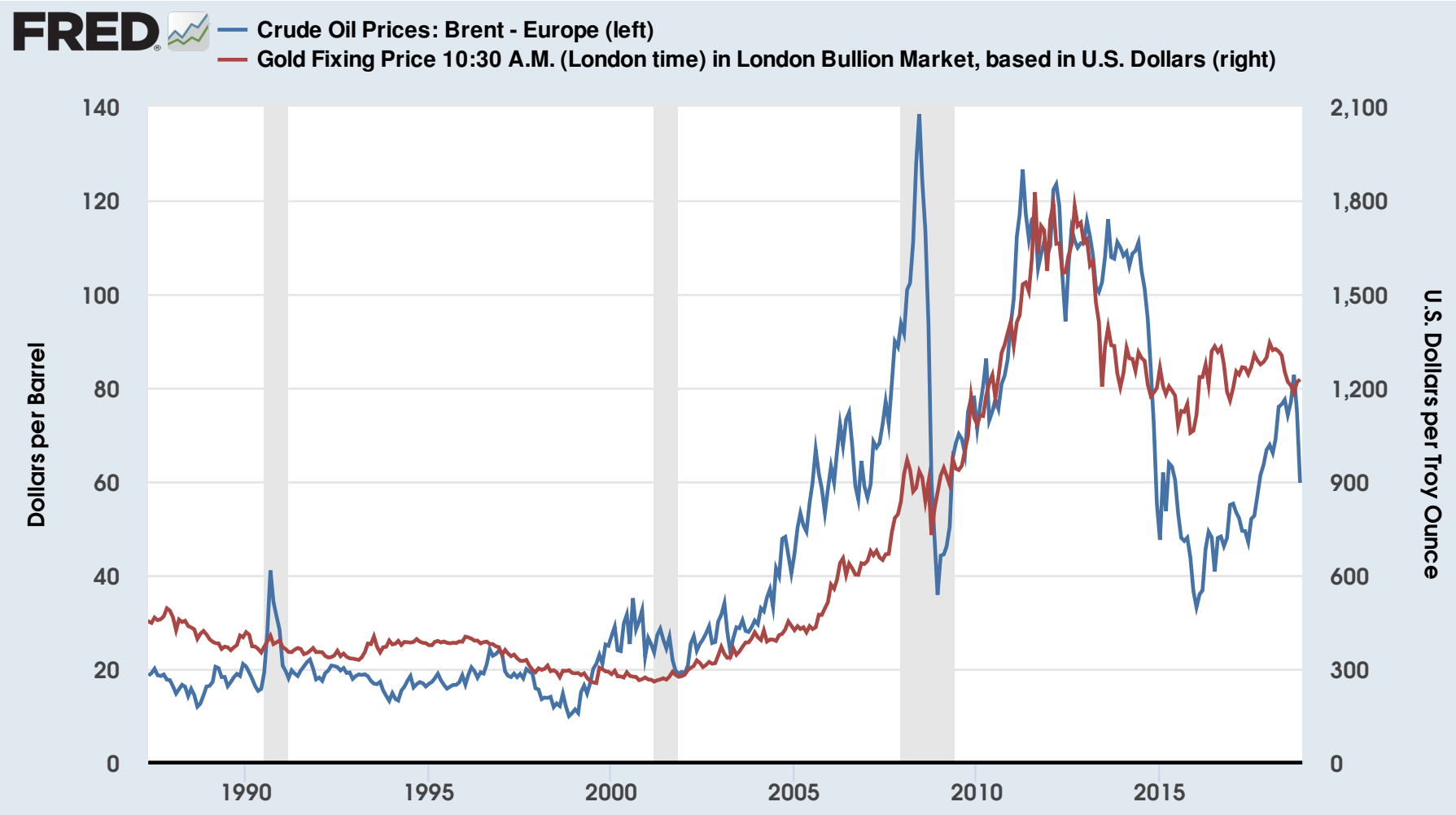
[a] Russell (1912), via Deaton and Cartwright (2018).

Causal Inference

Causal inference is concerned with the following questions:

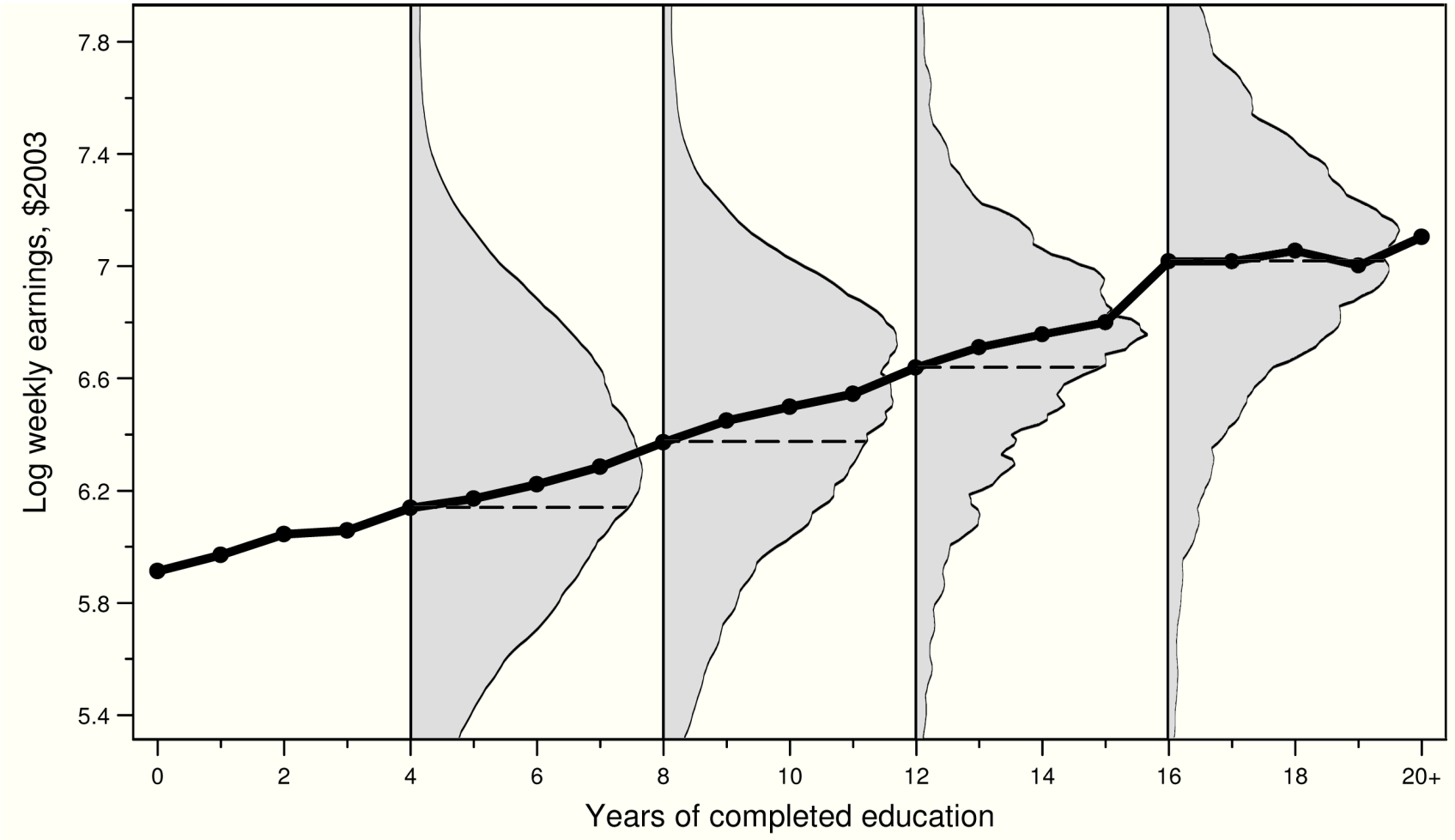
1. Does x have a causal effect on y? If so, how large is the effect? (**causal effect learning**)
2. If a causal effect exists, what is the mechanism by which it occurs? (**causal mechanism learning**)

Causal Inference



Do oil and gold prices cause each other to move?

Causal Inference



Does receiving more education make you earn more?

Program Evaluation

Evaluating and predicting the effects of government programs and economic policies is a central problem in applied economic research:

- Effect of worker training programs on employment
- Effect of early childhood interventions on adult outcomes
- Effect of negative income taxes on labor supply
- Effect of environmental regulations on pollution emission
- ...

Artificial Intelligence

- So far, progress in causal inference has been made mainly in developing methods to learn causal effects or estimate causal models from data *based on* our understanding of the underlying mechanisms.
- Models of causal mechanisms are developed by human experts.
  - Science progresses by formulating models of causal mechanisms, then conduct experiments or observational studies, and update the models based on their results.
- Building machines that can learn causal mechanisms without human experts would be the ultimate goal of artificial intelligence.

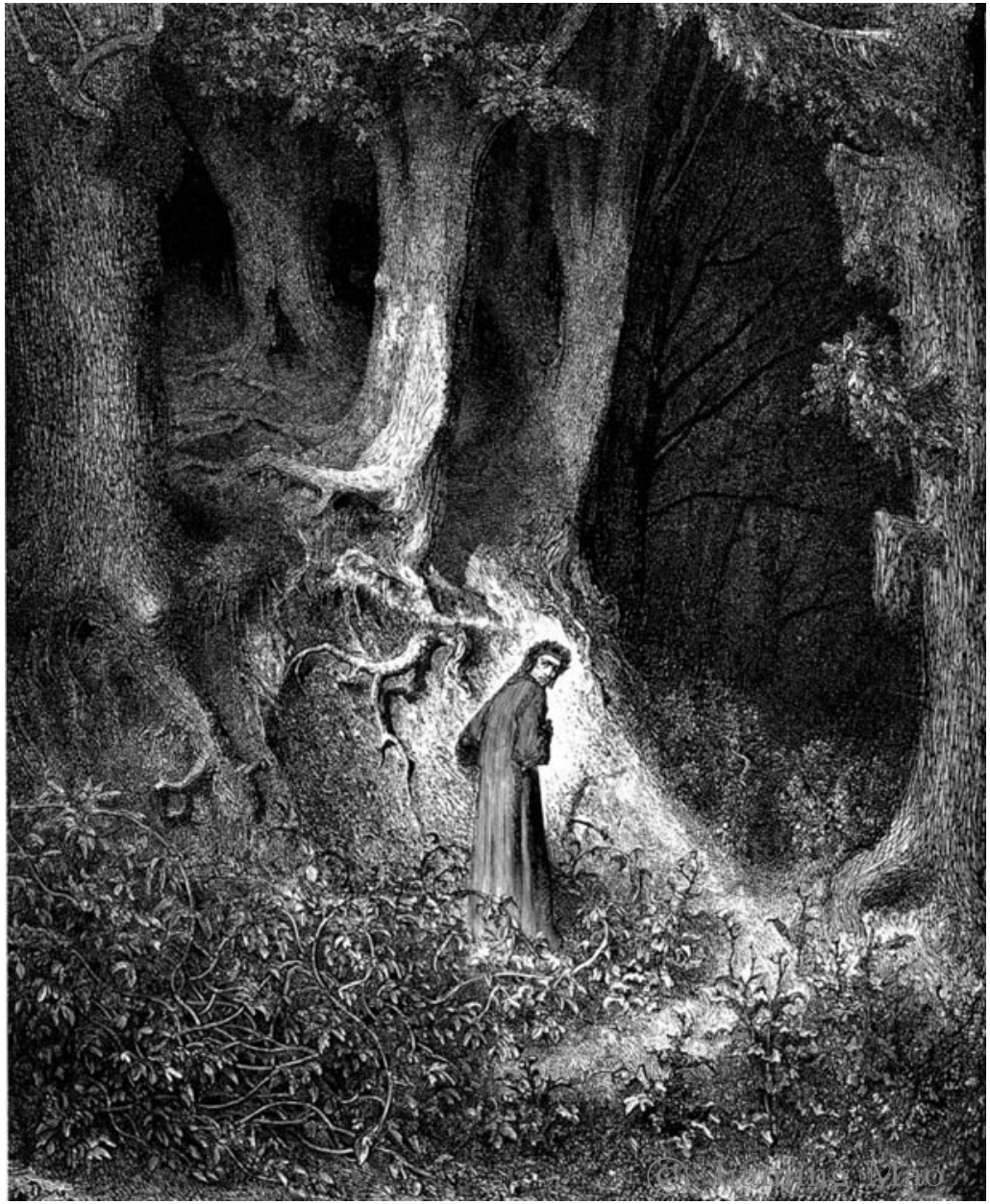
Road Map

1. Statistical Learning

2. Causal Inference

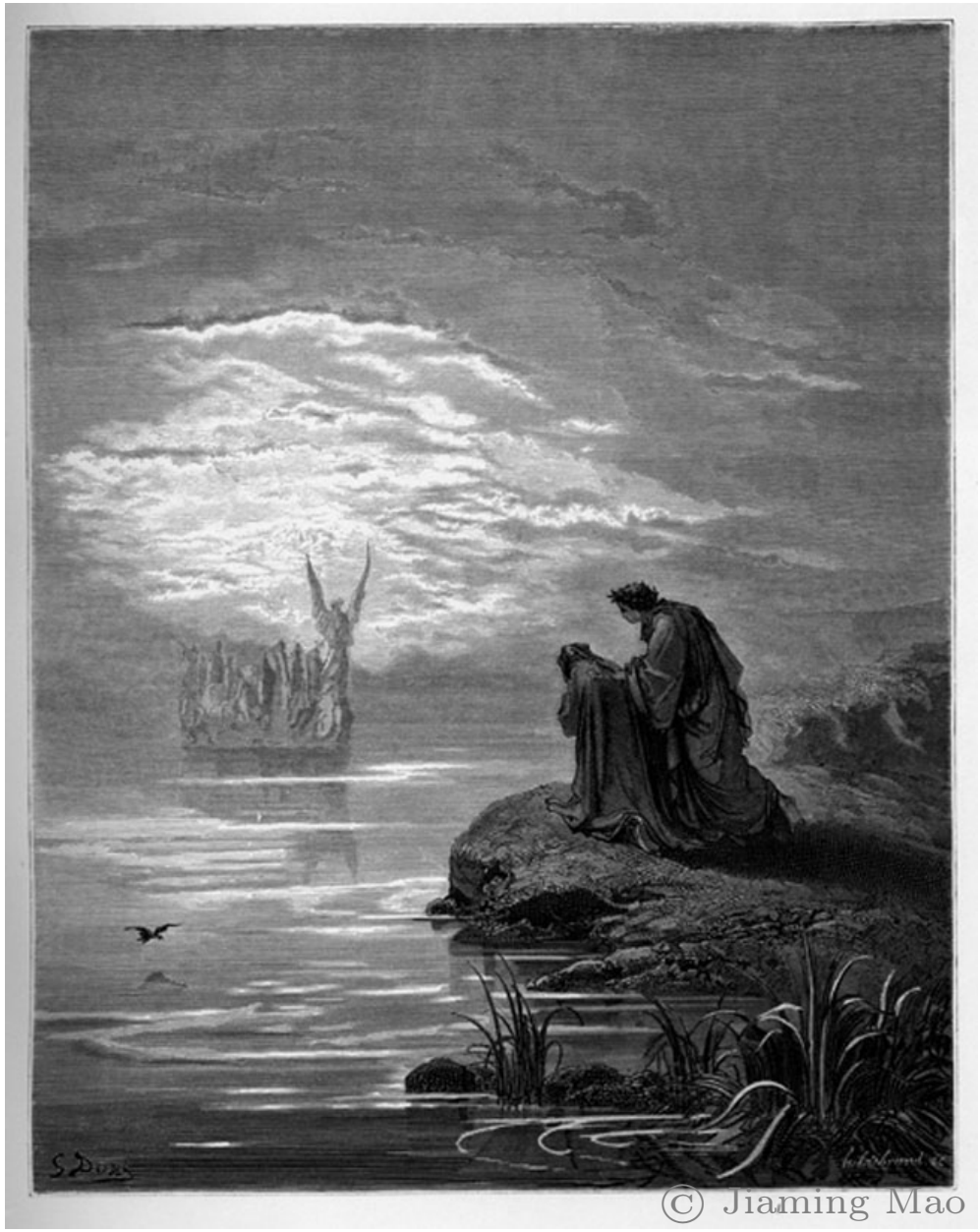
Road Map

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



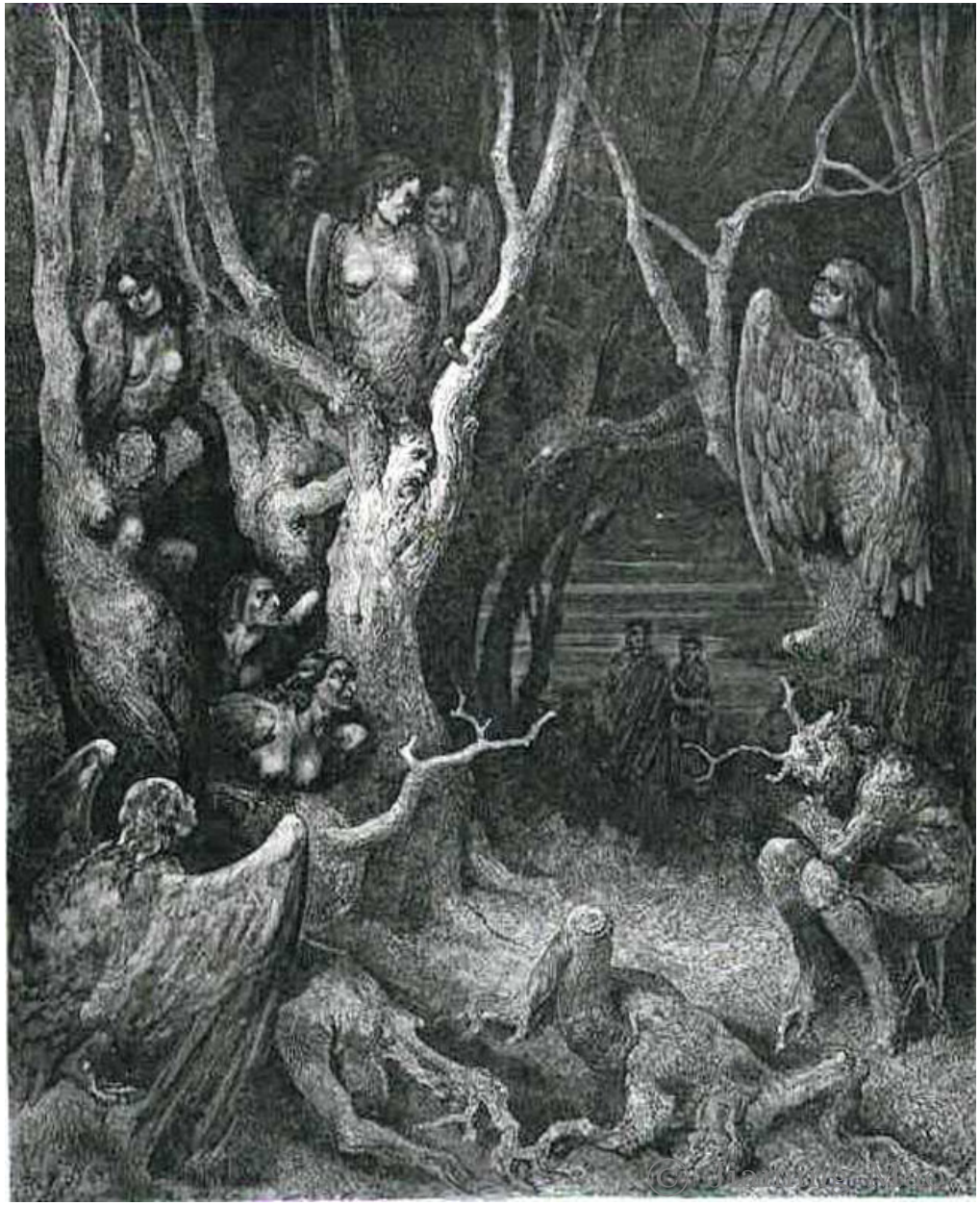
Road Map

... 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[1]
- 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.

[1] In this lecture, we use $p(x)$ to both denote the probability mass function (pmf) if x is a discrete random variable and the probability density function (pdf) if x is a continuous random variable.

Statistical Learning

Let

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

, where e is an error term.

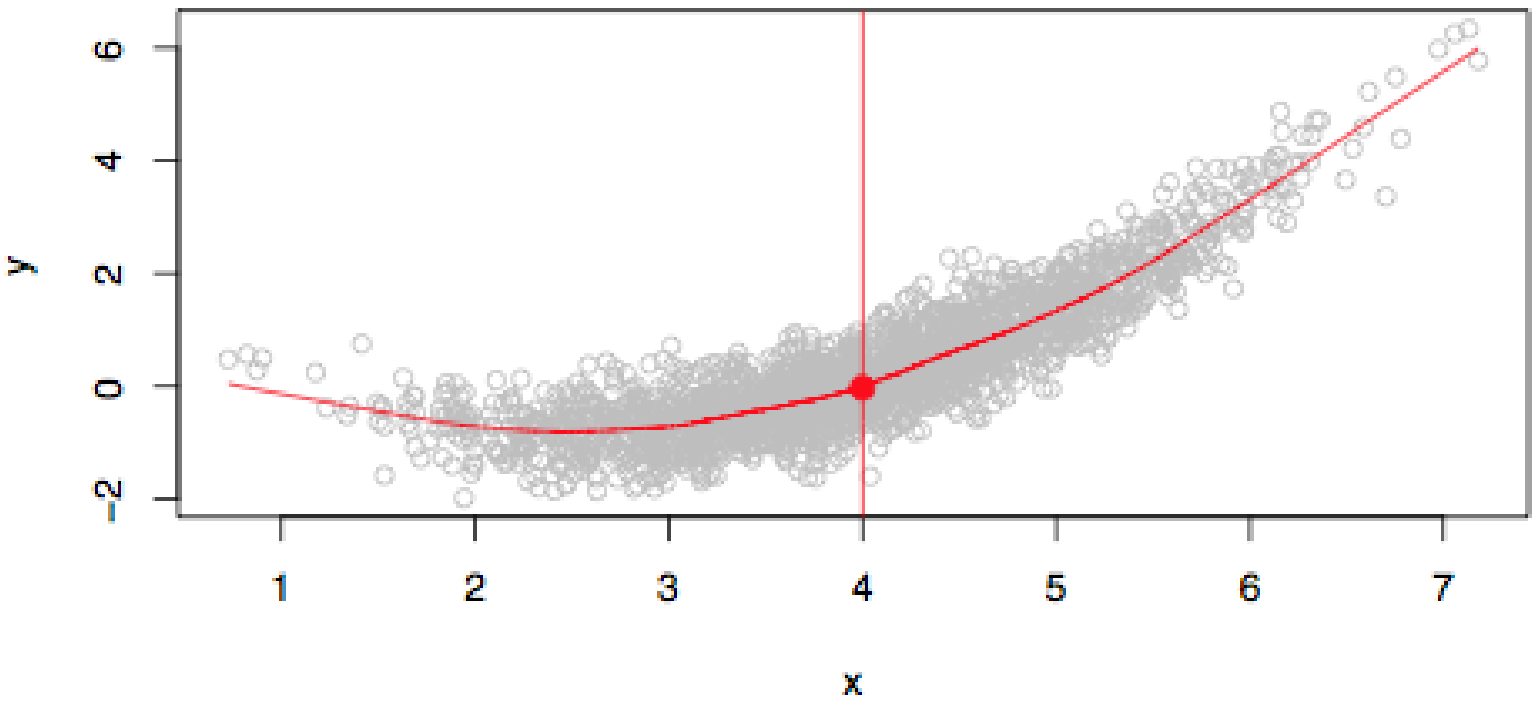
What is the function f that produces the best prediction of y given x?

- Depends on how we measure "best." Common choice: minimizing the expected squared-error loss[2] $E\left[(y - f(x))^2\right] \Rightarrow f(x) = E[y|x]$.
- $f(x) = E[y|x]$ is the **target function** that we want to learn[3].

[2] Also commonly called the **mean squared error (MSE)**.

[3] **Learning** is also called **estimation**. We will use the two terms interchangeably.

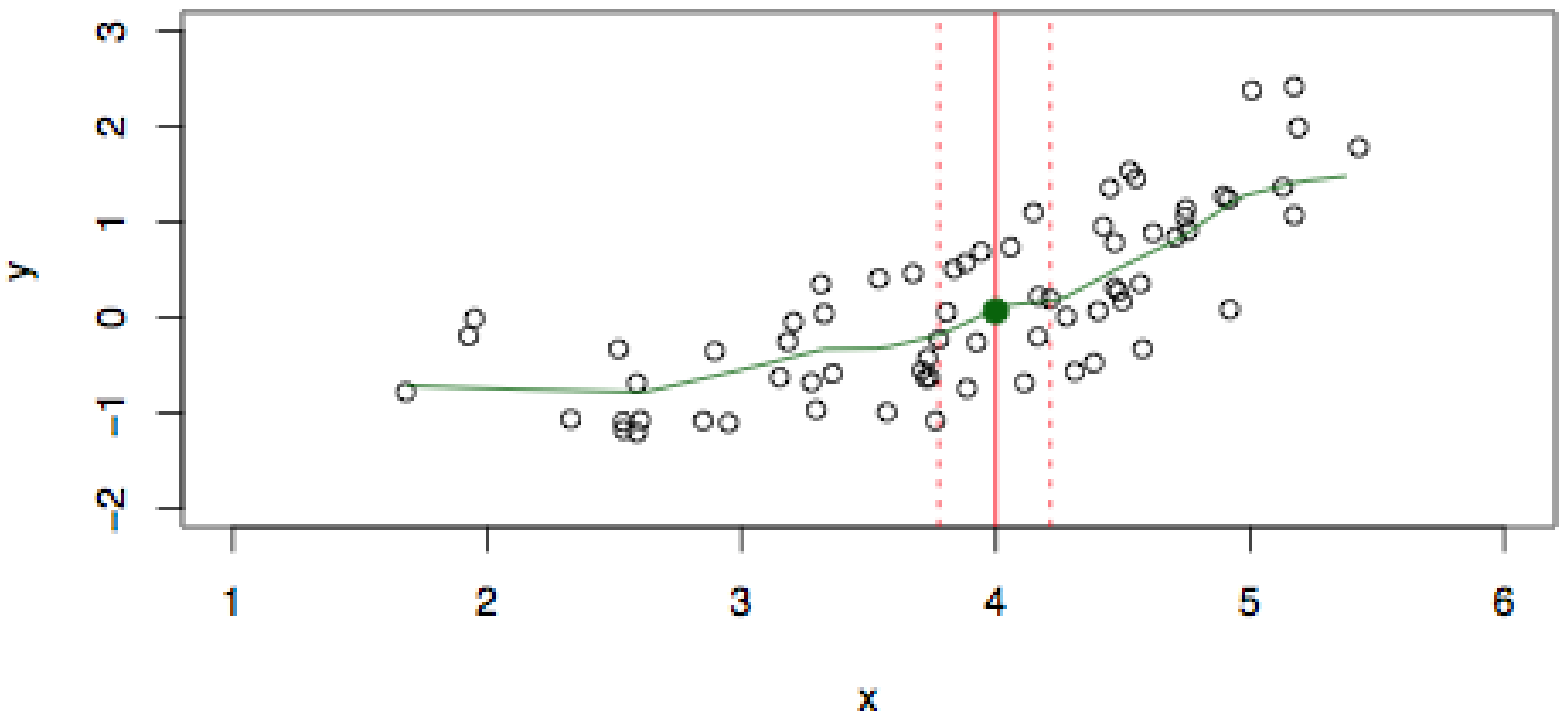
Learning f



$$\widehat{f}\,(x=4)=Ave\,(y|x=4)$$

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.



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

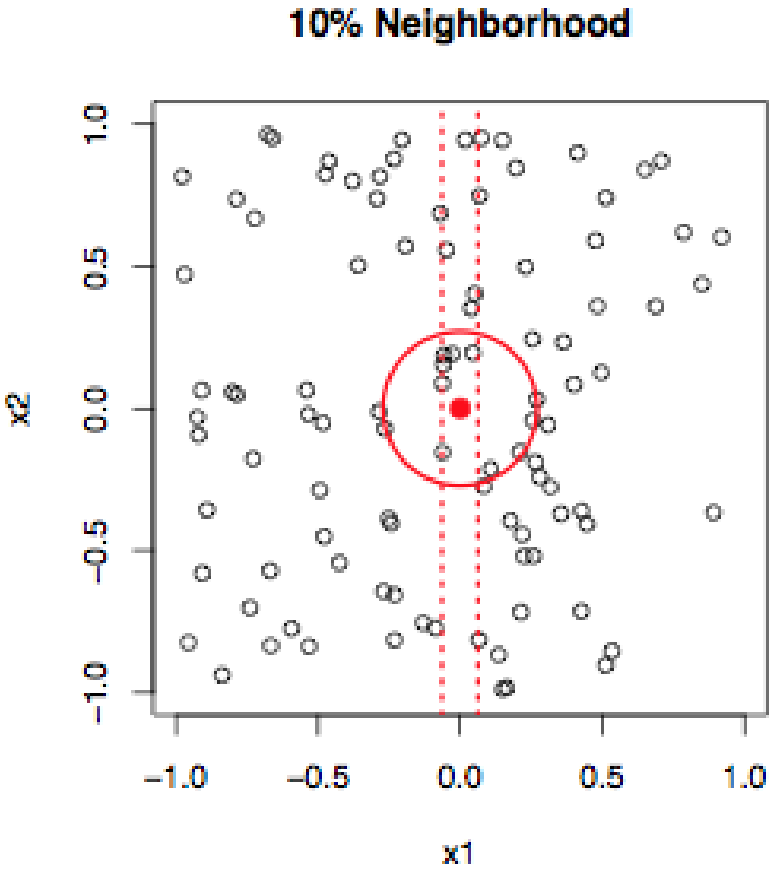
, where $\mathcal{N}\,(x)$ is some neighborhood of x.

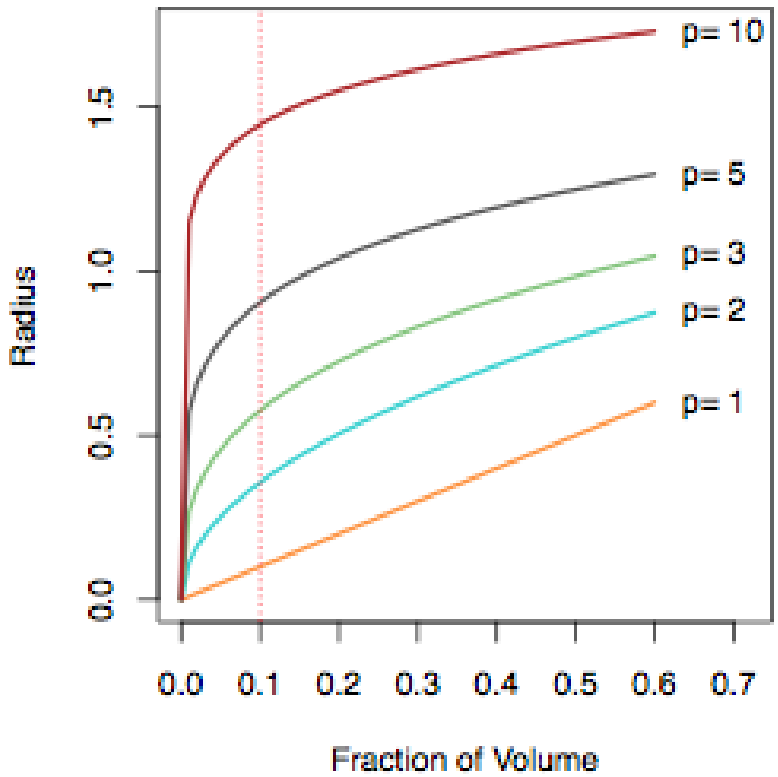
Learning f

- When x is multi-dimensional, i.e. $x = (x_1, \ldots, x_p)$, nearest neighbor averaging can work well for small p and large N[4].
- 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.**

[4] N : the number of data points

Learning f





Nearest neighbor and the curse of dimensionality

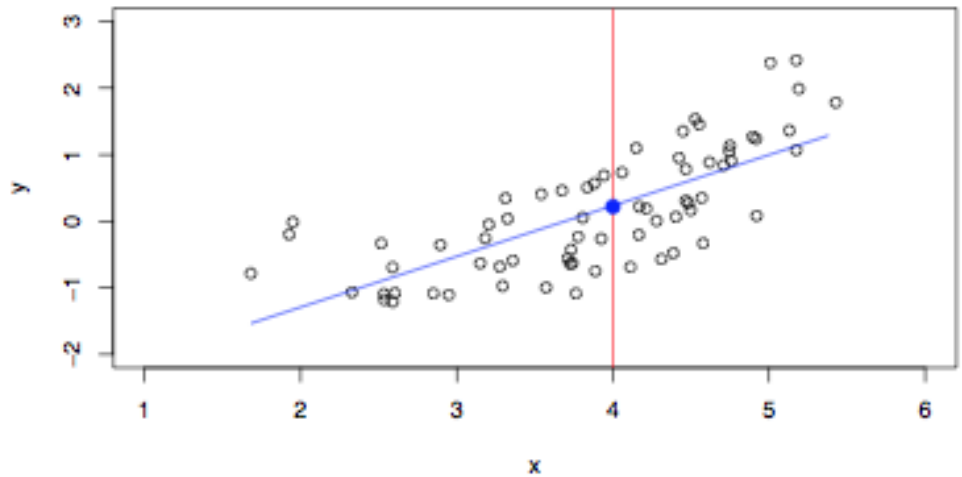
Learning f

- **Parametric methods**[5] of estimating $f(x)$ assume a specific functional form with a fixed number of parameters.
  - Linear regression: $f(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p = \beta' x$
- **Nonparametric methods** do not make explicit assumptions about the functional form of $f(x)$[6].
  - Nearest neighbor averaging is a nonparametric method.

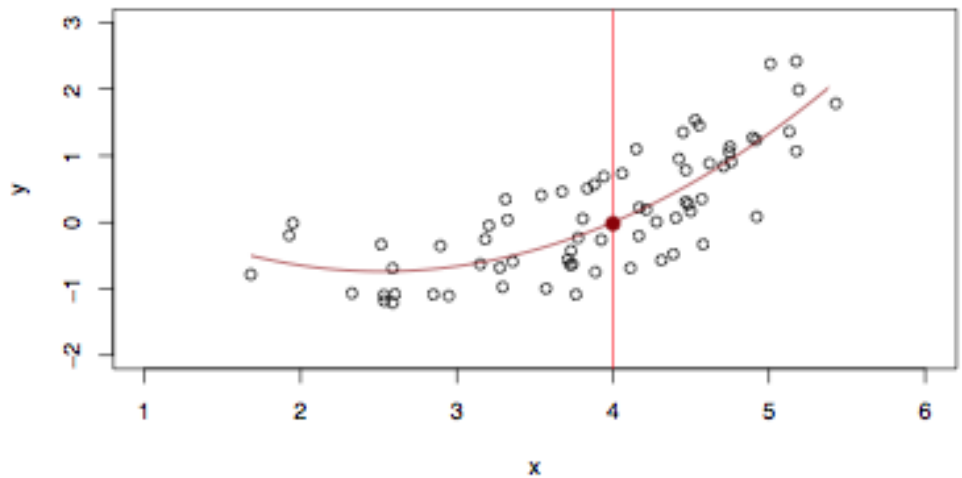
[5] We will use the terms "statistical method" and "statistical model" interchangeably.

[6] We will also learn methods that make some assumptions about the functional form of $f(x)$, but allow the number of parameters to grow with data. These methods are considered either *nonparametric* or *semiparametric*.

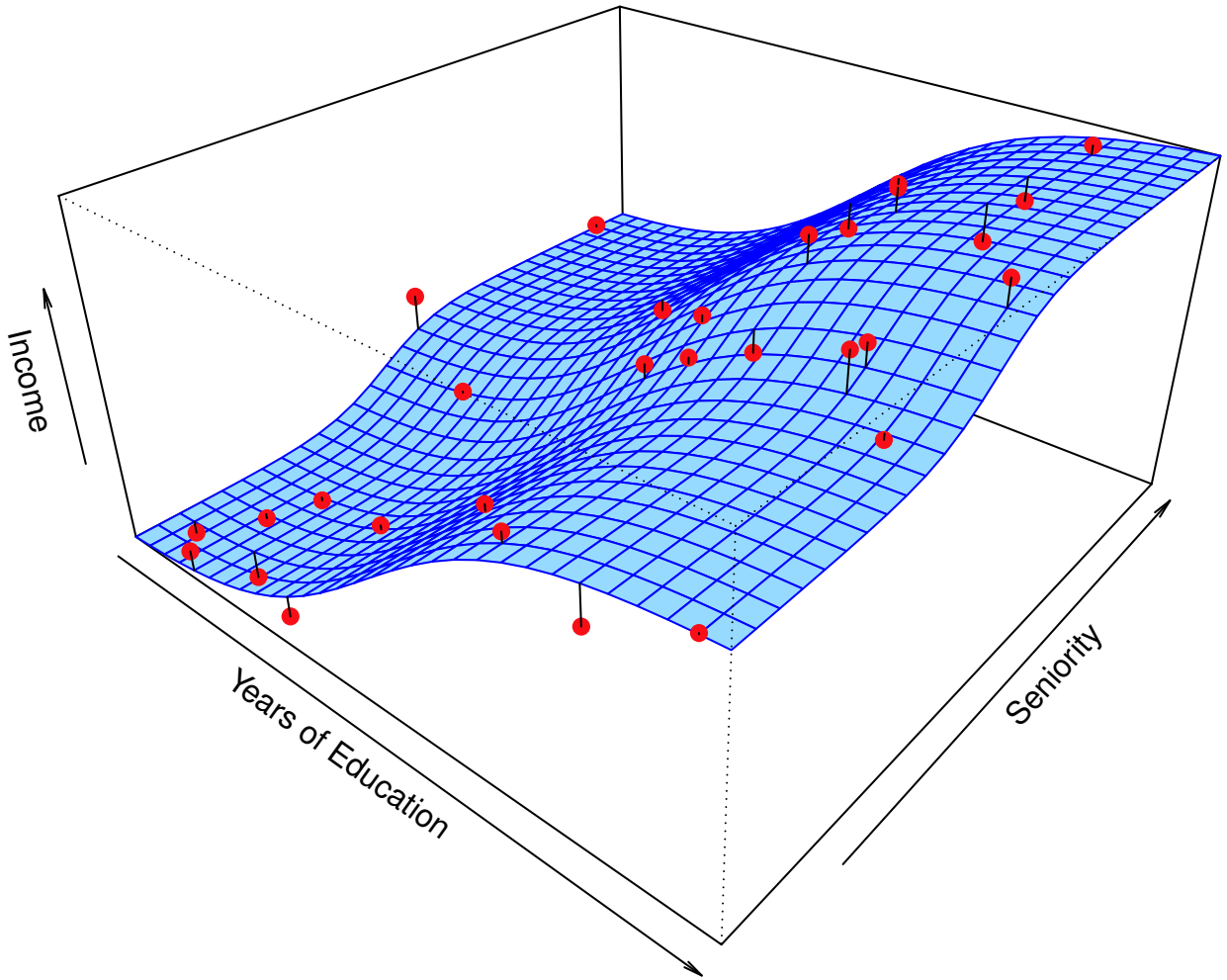
A linear model $\hat{f}_L(X) = \hat{\beta}_0 + \hat{\beta}_1 X$ gives a reasonable fit here



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

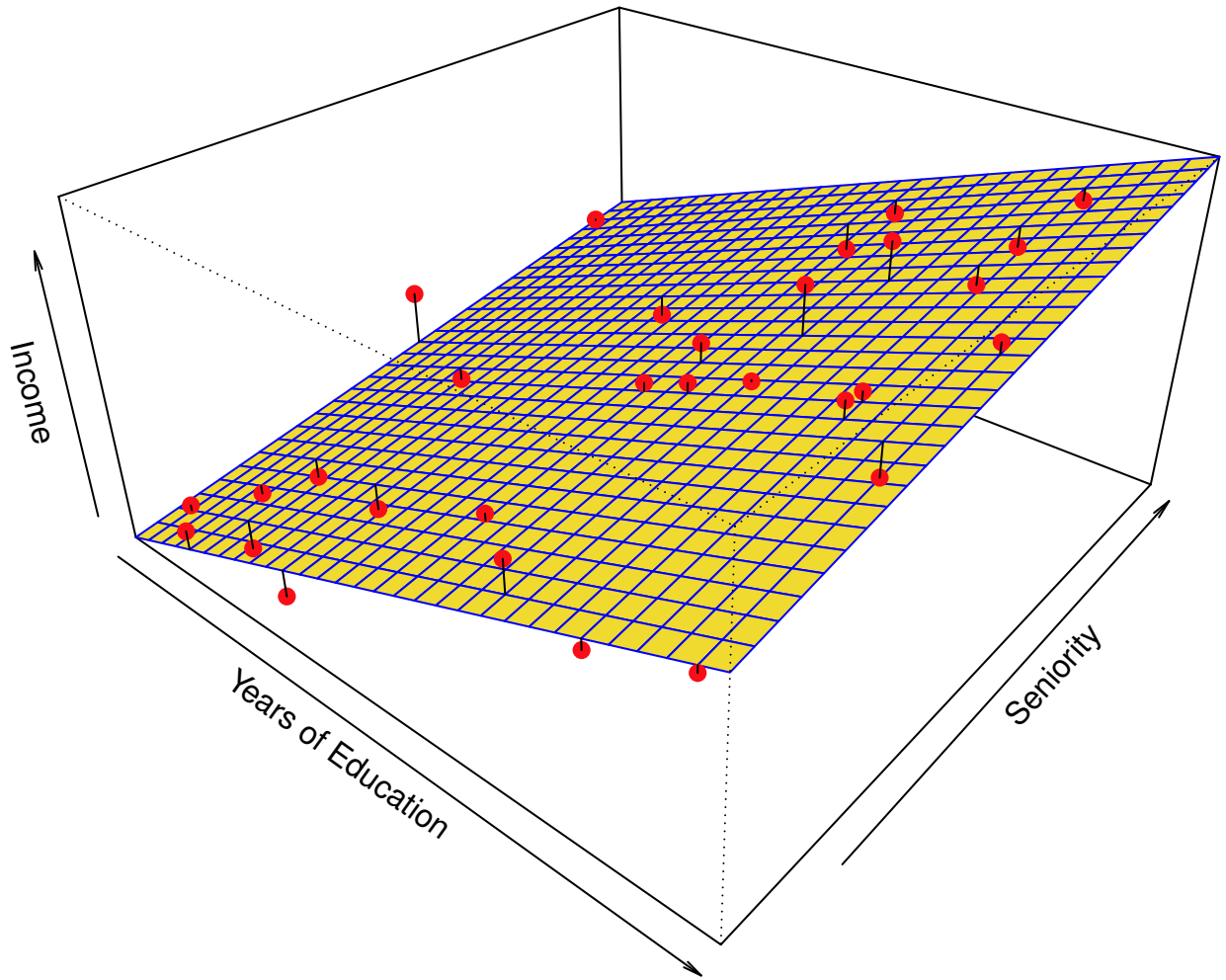


Learning f



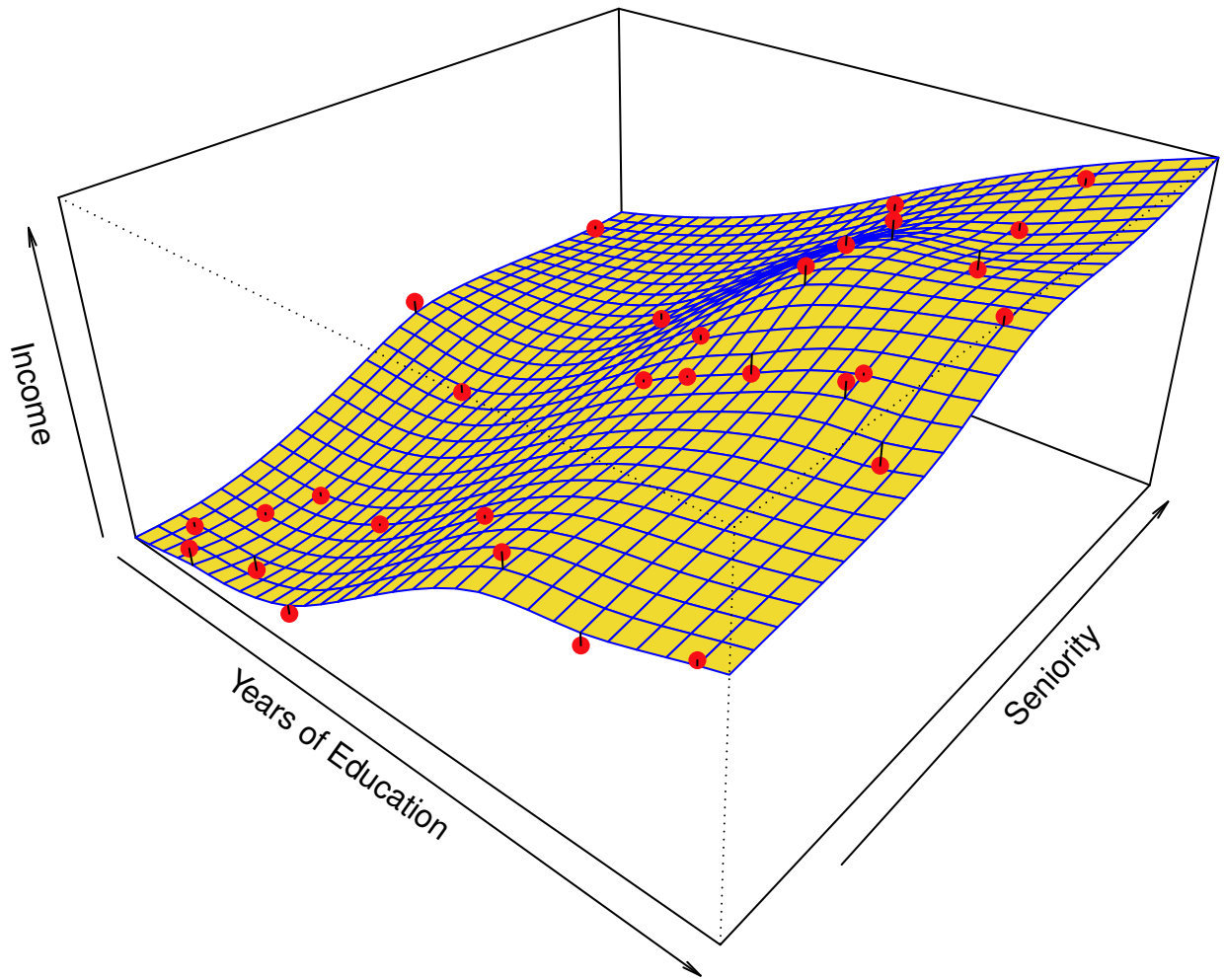
True f

Learning f



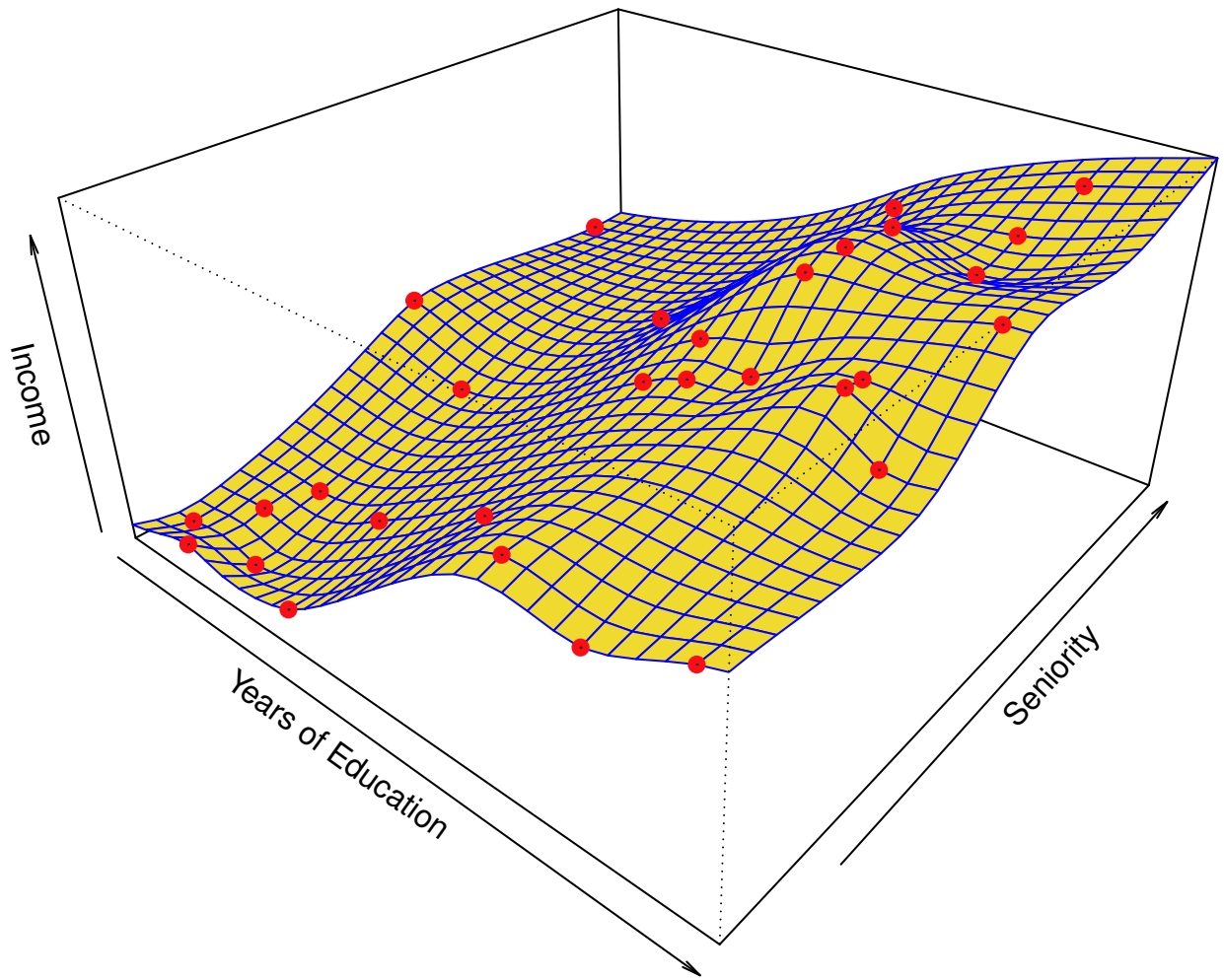
Linear Fit

Learning f



Thin-plate Spline Fit (Smooth)

Learning f



Thin-plate Spline Fit (Rough)

Here $\widehat{f}$ fits the data perfectly: $\widehat{f}(x)$ contains not only $f(x)$ but also e.

Assessing the Goodness of Fit

Let $\mathcal{D}_{TR} = \{(x_1, y_1), \ldots, (x_N, y_N)\}$ denote the data on which we estimate f. This is called **training data**.

We can assess how well $\widehat{f}$ fits the training data by calculating the **training error**:

$$\text{error}_{TR} = \frac{1}{N} \sum_{i \in \mathcal{D}_{TR}} \left(y_i - \widehat{f}(x_i) \right)^2$$

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

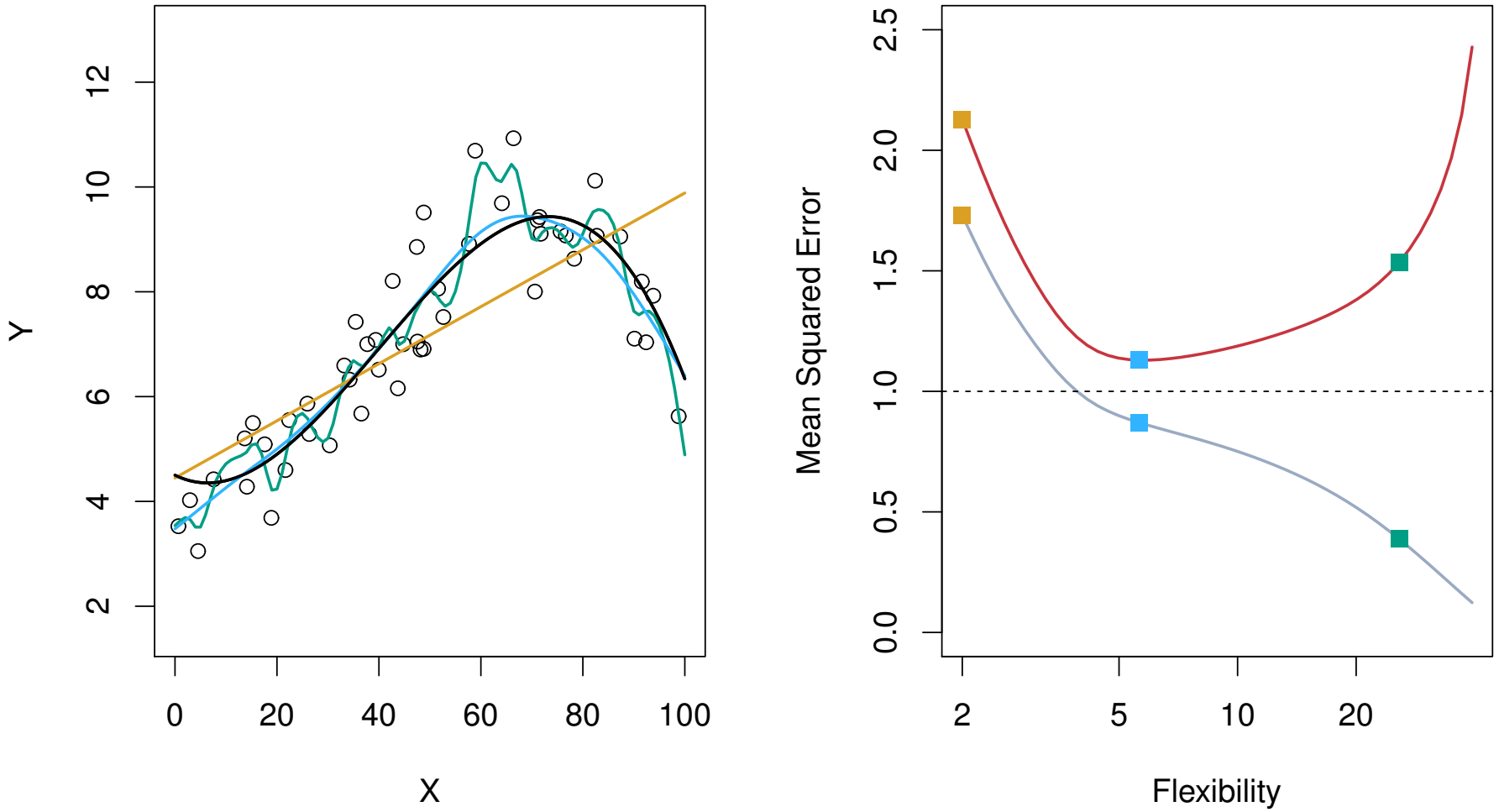
Assessing the Goodness of Fit

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

$$\text{error}_{TE} = \frac{1}{M} \sum_{i \in \mathcal{D}_{TE}} \left(y_i - \widehat{f}(x_i) \right)^2$$

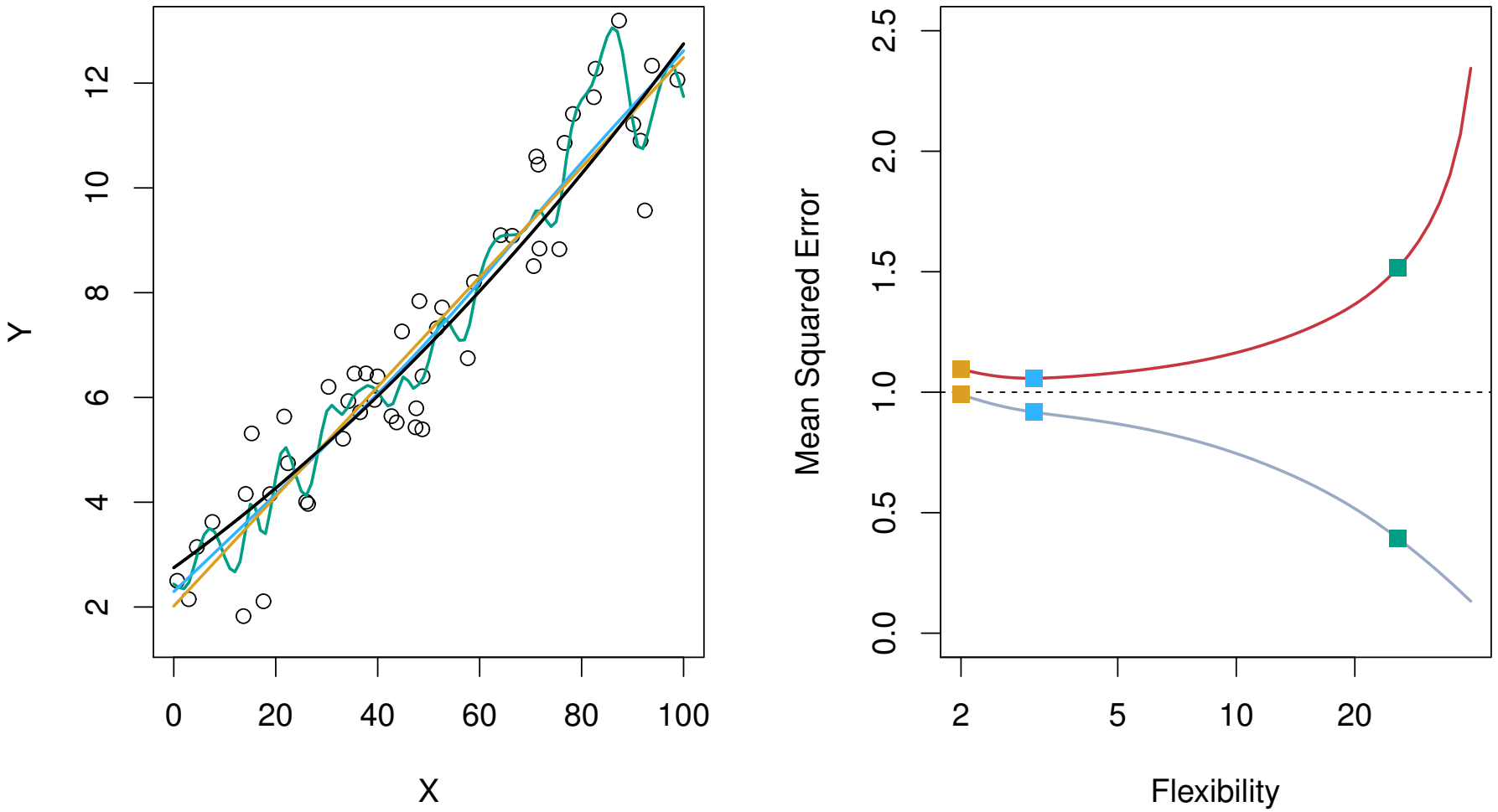
When $M \to \infty$, $\text{error}_{TE} \to \underbrace{E\left[\left(y - \widehat{f}(x) \right)^2 \right]}_{\textbf{prediction error}}$.

Assessing the Goodness of Fit



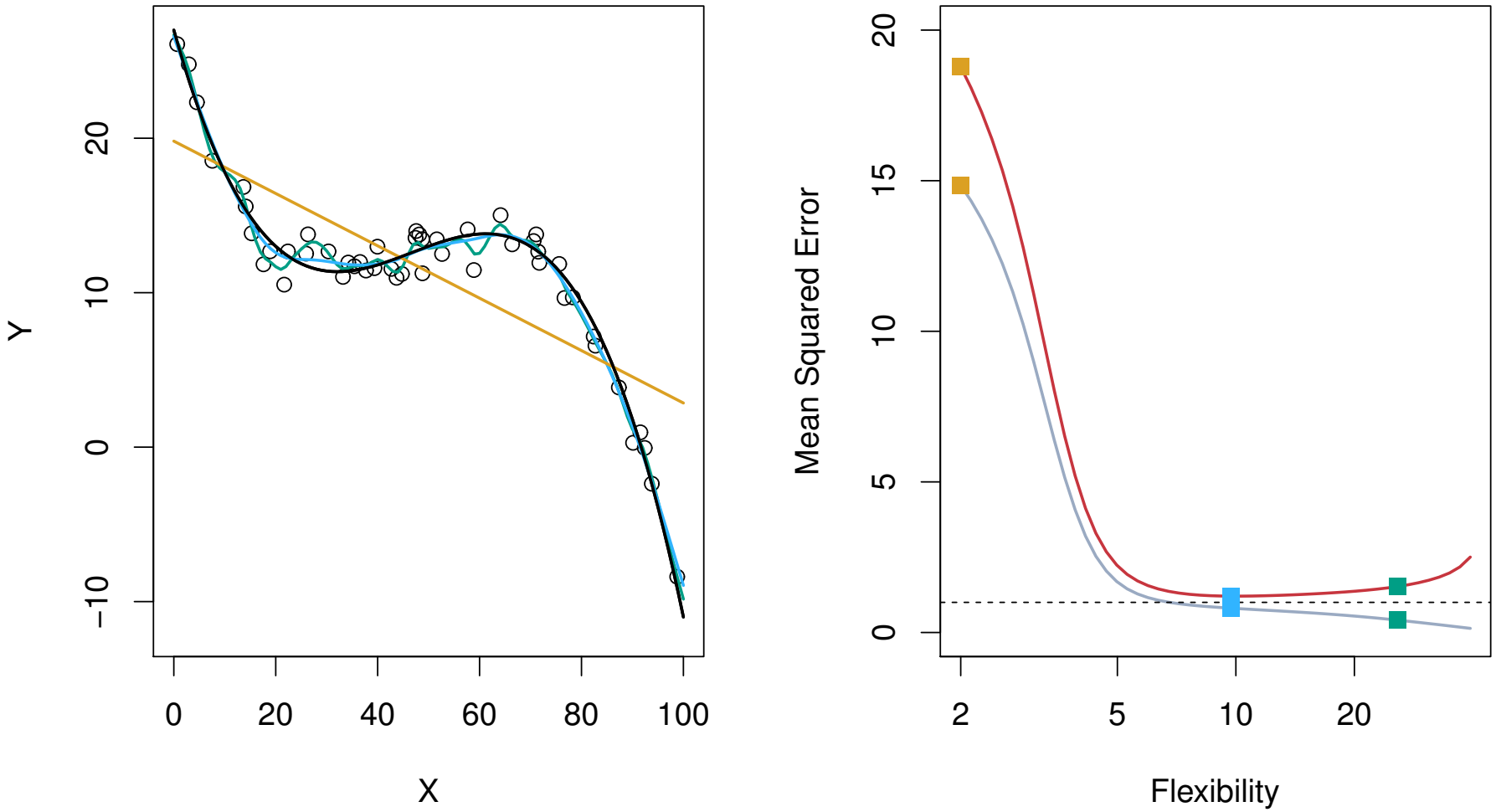
Left: true f (black), linear fit (orange), smoothing spline fits (blue & green).
Right: training error (grey), prediction error (red), Var(e) (dashed).

Assessing the Goodness of Fit



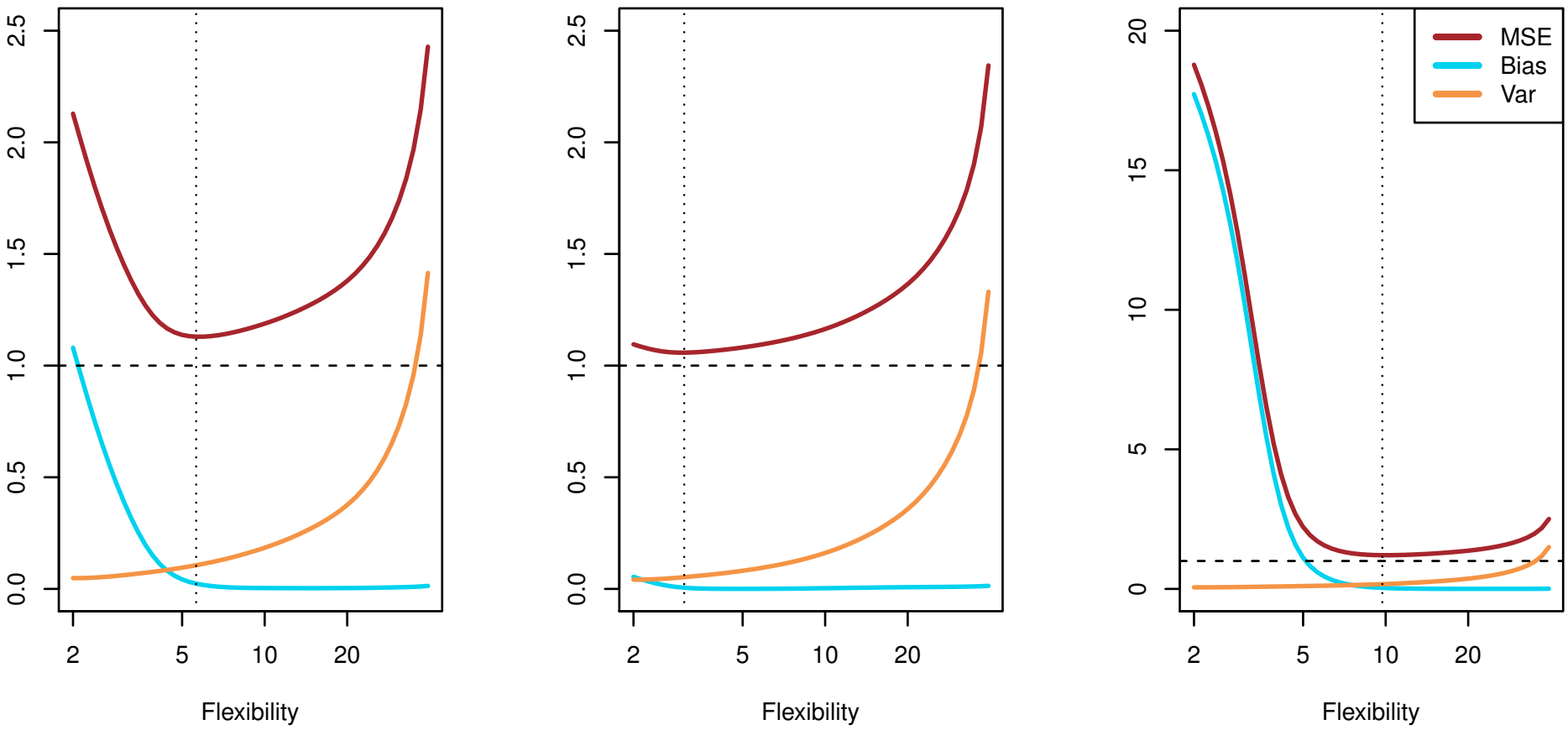
Left: true f (black), linear fit (orange), smoothing spline fits (blue & green).
Right: training error (grey), prediction error (red), Var(e) (dashed).

Assessing the Goodness of Fit



Left: true f (black), linear fit (orange), smoothing spline fits (blue & green).
Right: training error (grey), prediction error (red), 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 $\widehat{f}$ of f and $\widehat{y} = \widehat{f}(x)$,

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

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

The Bias-Variance Trade-off

- $Var\left(\widehat{f}\right)$ refers to the amount by which $\widehat{f}$ would change if we estimate it using a different training data set.

- As a general rule, as model flexibility increases, bias $\left(\widehat{f}\right)$ will decrease and $Var\left(\widehat{f}\right)$ will increase.

  - More flexible models tend to have higher variance because they have the capacity to follow the data more closely. Thus changing any of the data points may cause the estimate $\widehat{f}$ to change considerably.

The Bias-Variance Trade-off

- As the flexibility of the model increases, we observe a monotone decrease in training error and a U-shape in prediction error.

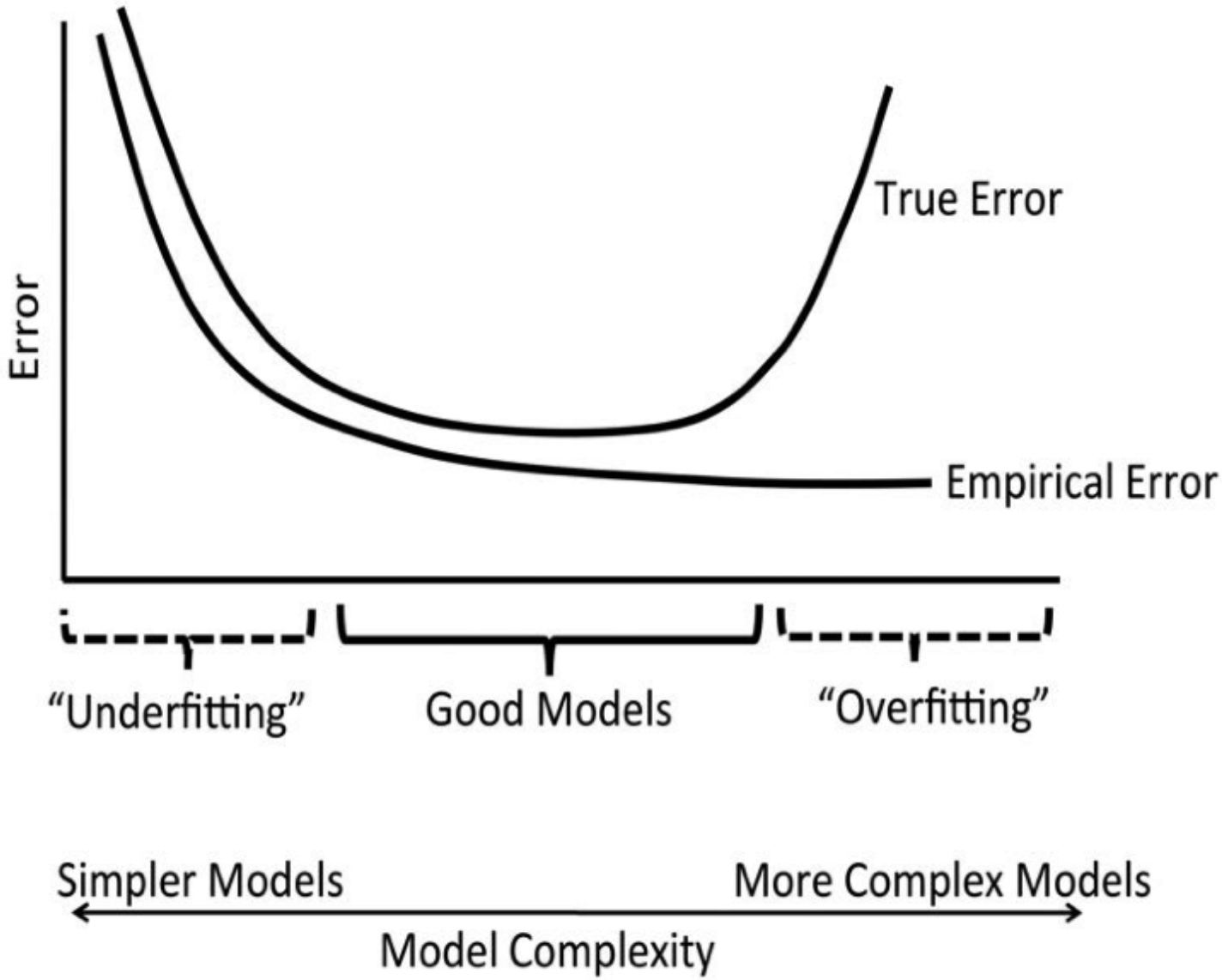
- This is due to the **bias-variance trade-off**: as model flexibility increases, the bias tends to initially decrease faster than the variance increases. Then at some point increasing flexibility has little impact on the bias but starts to significantly increase the variance.

The Bias-Variance Trade-off

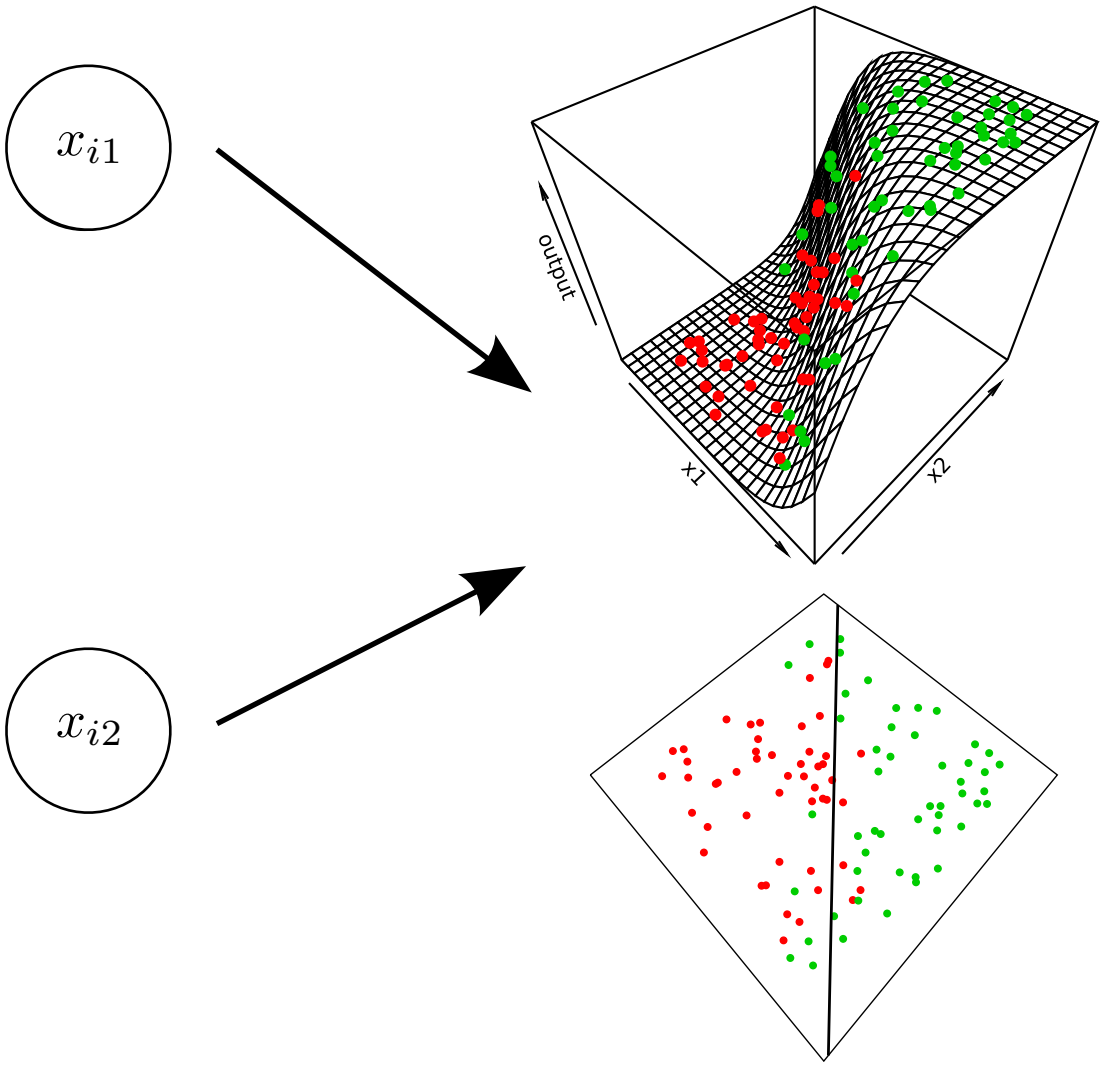
- The bias-variance trade-off is a trade-off because it is easy to have a model with extremely low bias but high variance (e.g., by drawing a curve that passes through every single training observation) or one with very low variance but high bias (e.g., by fitting a horizontal line to the data). The challenge lies in finding a model 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.

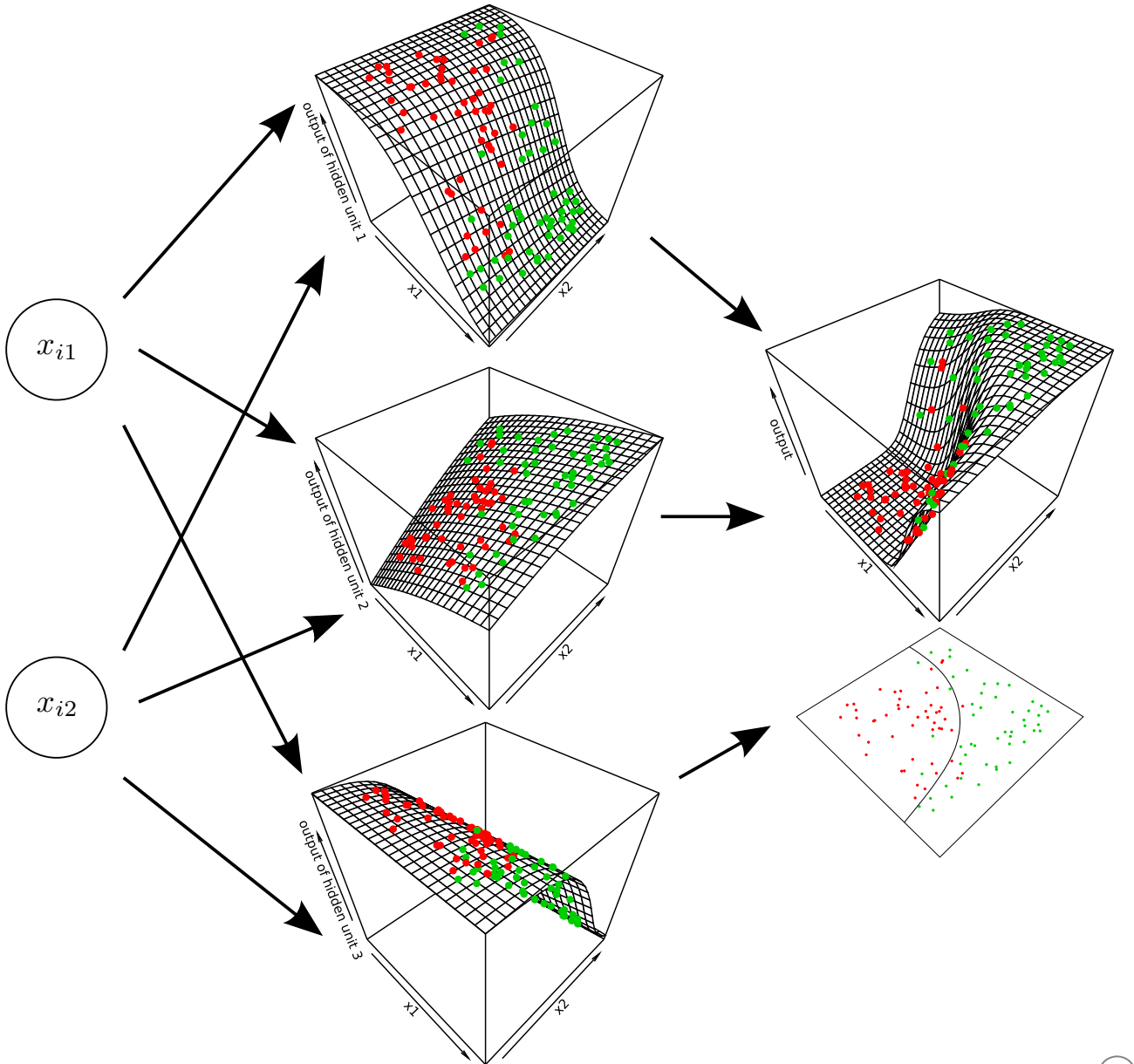
The Bias-Variance Trade-off



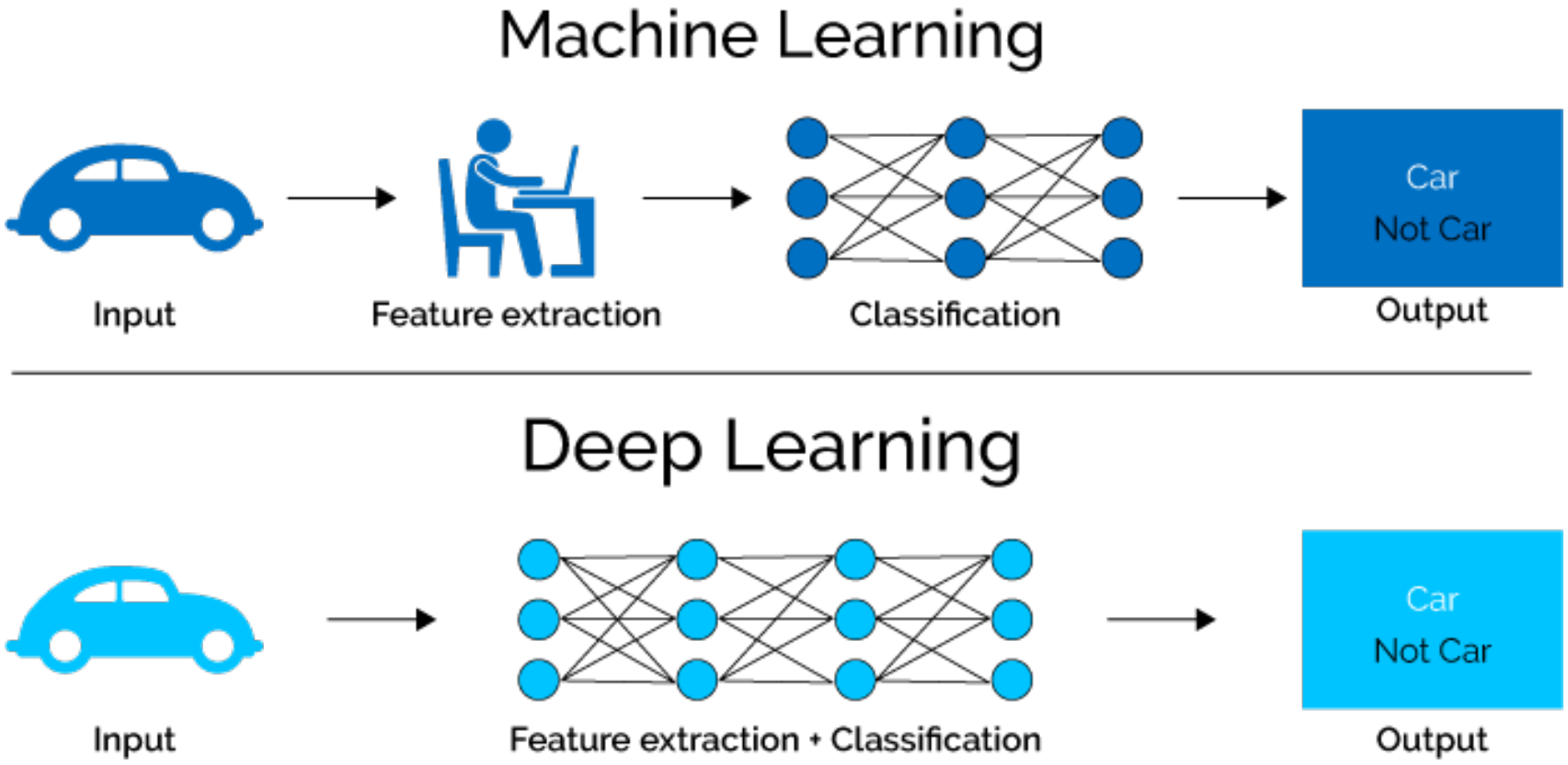
From Shallow to Deep Learning



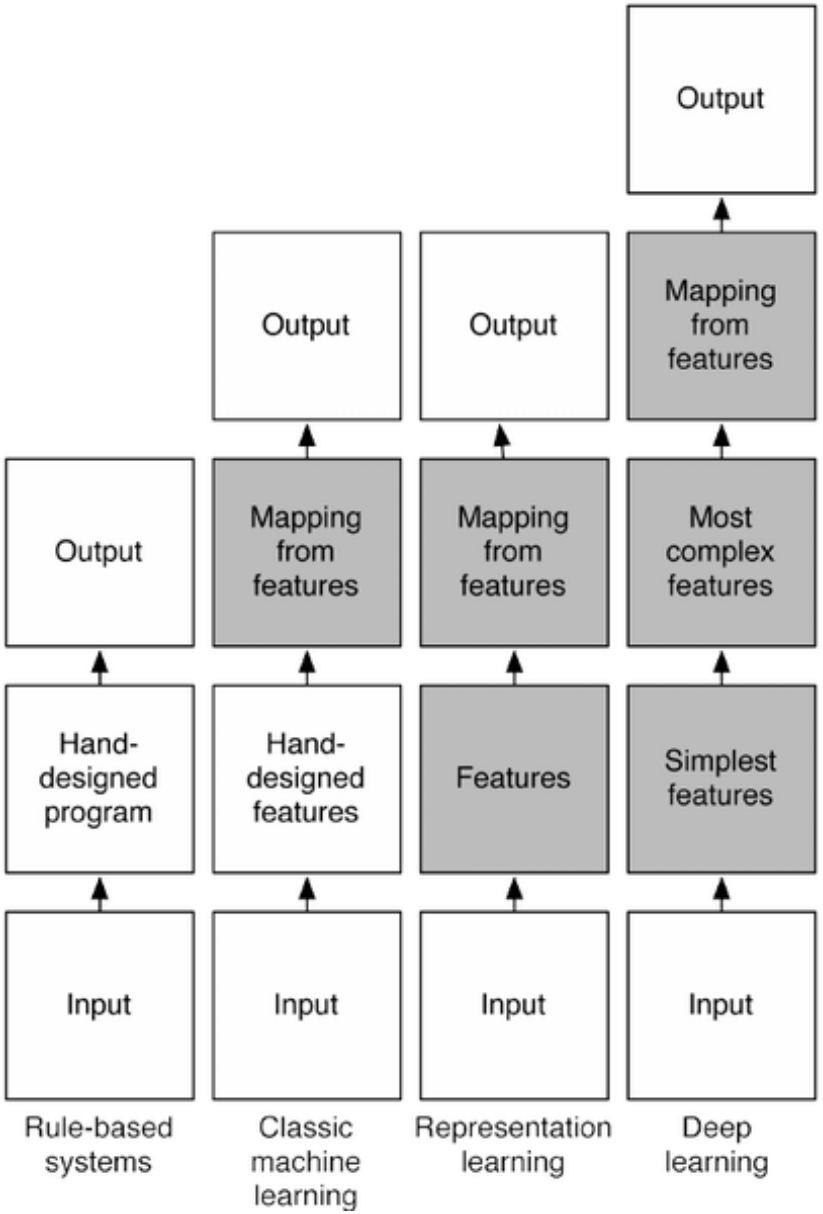
From Shallow to Deep Learning



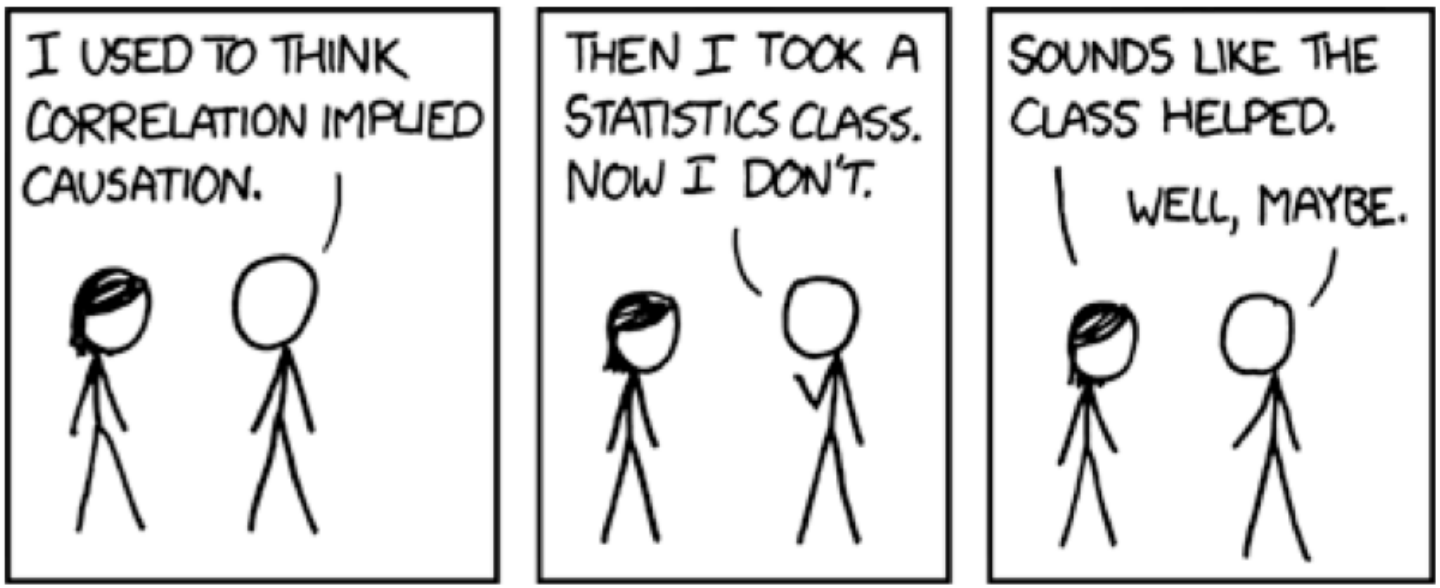
From Shallow to Deep Learning



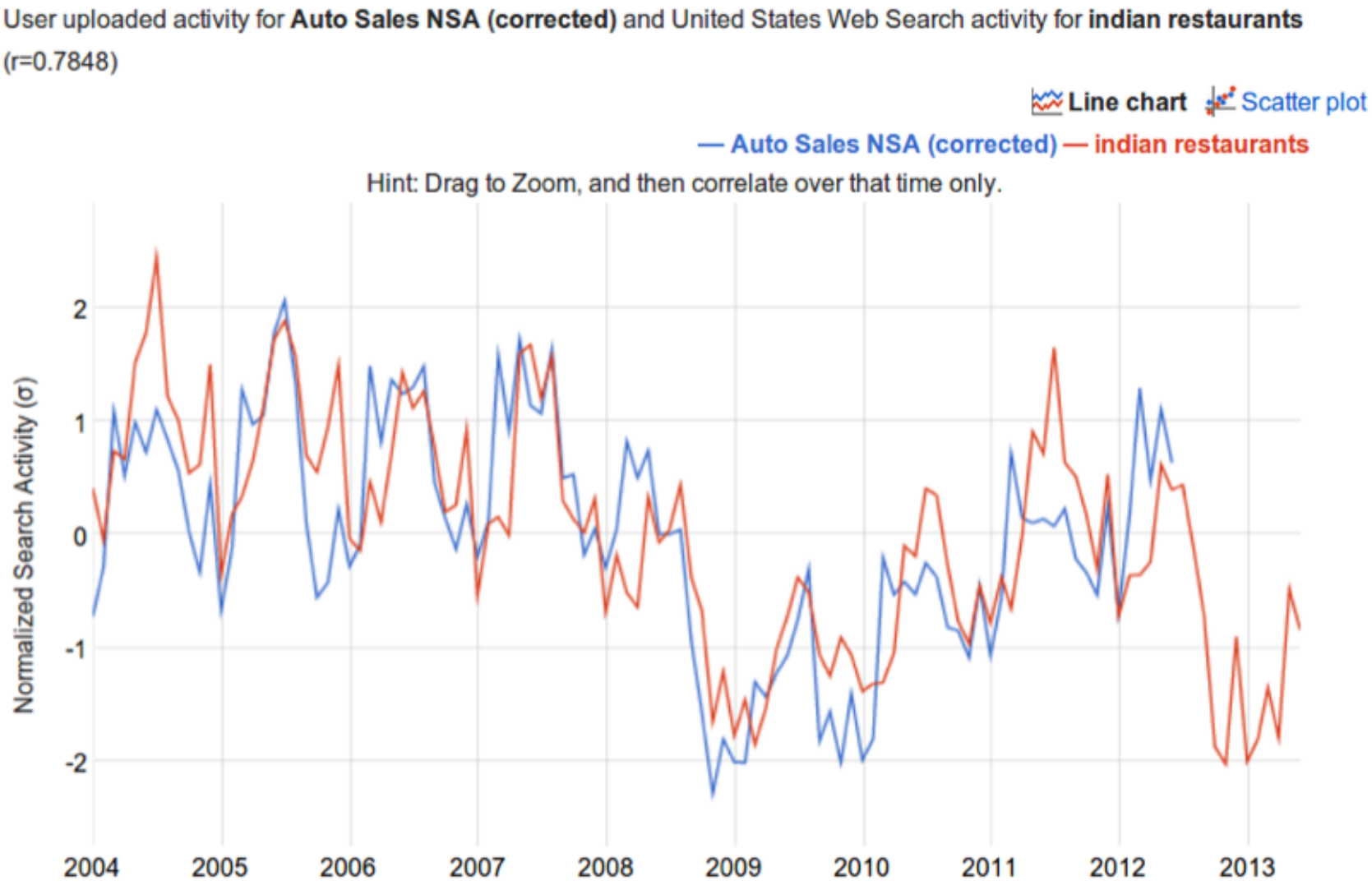
From Shallow to Deep Learning



Causal Inference



Correlation does not imply Causation



Auto Sales and Search for Indian Restaurants. Source: Google Correlate

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[7] 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*."

[7] 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* $x = a$?
  - $E[y|\text{do}(x = a)]$[8]
- **Statistical prediction**: What will y be if I *observe* $x = a$?
  - $E[y|x = a]$

[8] Assuming we minimize the expected L2 loss in prediction.

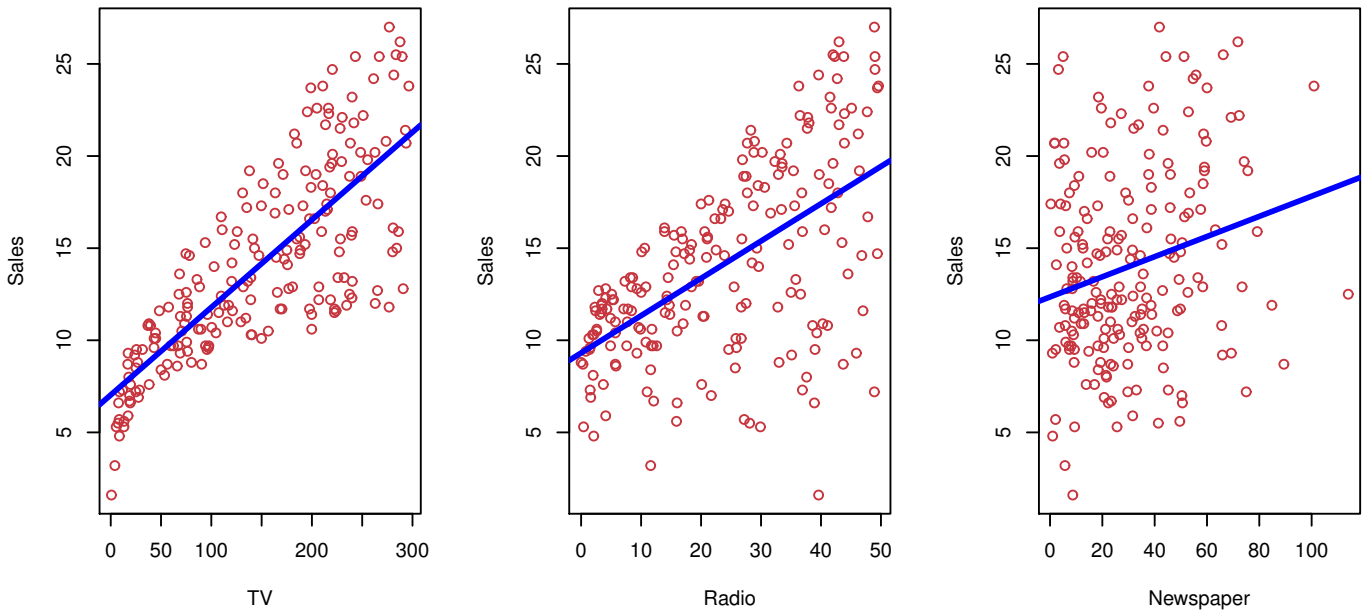
Hospitalization and Health

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

- To learn $f(x) = E[y|\text{do}(x)]$, 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

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

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

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

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

For data generated by randomized experiments, correlation implies 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

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

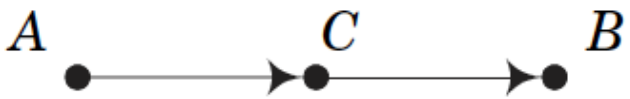
- For observational data, correlation no longer implies causation. Consider the example of hospitalization and health: the fact that hospitalization is associated with worse health outcomes may not be due to the adverse effect of hospitalization on health, but to the fact the people with worse health 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 Diagrams

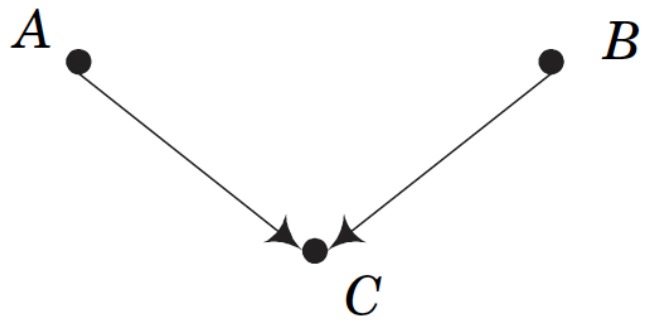
- To learn causal effects from observational data, we need to have an understanding of the **causal mechanism** that generates the data[9].
- **Causal diagrams** are graphs that can be used to represent causal relationships and therefore describe our qualitative knowledge about a causal mechanism.

[9] 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!

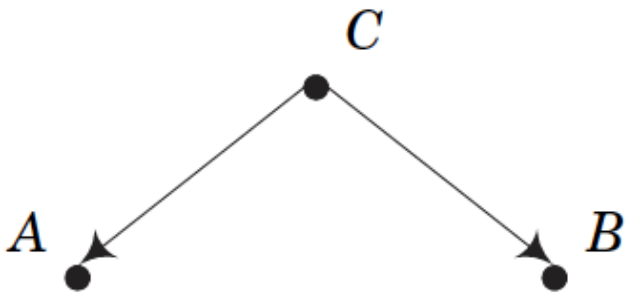
Causal Diagrams



(a) Mediation



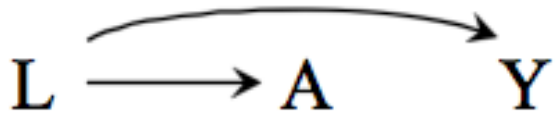
(c) Mutual causation



(b) Mutual dependence

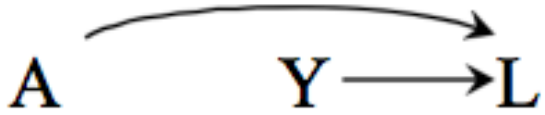
Basic patterns of causal relationships among three variables

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 associated: 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



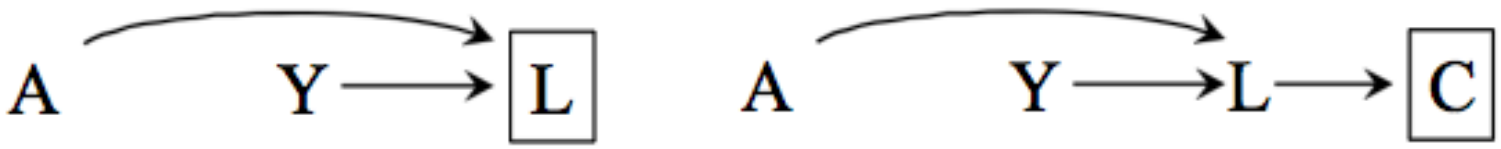
- Both A and Y have a causal effect on L. A does not have a causal effect on Y.
- L is called a **common effect** of A and Y.
- A and Y are independent.
- E.g., A : family heart disease history; Y : smoking; L : heart disease

Association and Causation

Box indicates conditioning

- A and Y are conditionally independent after conditioning on 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



- A and Y are conditionally associated after conditioning on 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[10]
2. They share common causes
3. The analysis is conditioned on their common effects[11]

[10] either directly or through mediating variables.

[11] 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.

Learning E[y|do(x)]

A basic strategy to learn the causal effect of x on y is to condition on their common causes[12] (while avoiding to condition on any of their common effects).

- Conditioning on common causes make two variables independent if they do not have direct causal effects on each other.
- Therefore, any association between two variables after their common causes have been conditioned on should be due to causation.

[12] When we condition on a variable, we also say we **control for** the variable.

Learning E[y|do(x)]

Let z be the set of common causes of x and y, and suppose any causal effect between x and y, if exists, flows from x to y, then

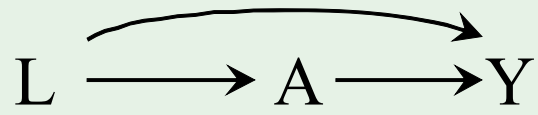
$$E\left[y|\text{do}\left(x\right)\right] = \sum_{z} E\left[y|\text{do}\left(x\right), z\right] p\left(z\right)$$
$$\overset{[1]}{=} \sum_{z} E\left[y|x, z\right] p\left(z\right)$$

, where [1] follows because, once z is conditioned on, association between x and y implies causation: $E\left[y|\text{do}\left(x\right), z\right] = E\left[y|x, z\right]$.

If we have data on $\{x, y, z\}$, then we can learn $E\left[y|x, z\right]$ and $p\left(z\right)$ using any statistical models of our choice, from which we obtain $E\left[y|\text{do}\left(x\right)\right]$.

Education and Earnings

Suppose the relationship between education and earnings is described by the following causal diagram:



L : ability; A : education; Y : income

Then

$$E\left[Y|\text{do}\left(A\right)\right]=\sum_{L}E\left[Y|A,L\right]p\left(L\right)$$

If we have data on education, income, and ability (e.g., by using test scores as proxy), then we can learn $E\left[Y|\text{do}\left(A\right)\right]$ by estimating $E\left[Y|A,L\right]$ and $p\left(L\right)$ from data[a].

---

[a] We can, for example, fit a linear model to $E\left[Y|A,L\right]$:

$$E\left[Y|A,L\right]=\beta_0+\beta_1 A+\beta_2 L$$

and use the empirical distribution for $p\left(L\right)$.

Structual Estimation

- **Causal models**, or **scientific models**, are mathematical models of causal mechanisms.
- In the econometrics literature, causal models based on economic theory are referred to as **structural models**. These models use economic theory to specify the functional forms of causal relationships.
- The estimation of structural models is called **structural estimation** – rather than learning a specific causal effect, structural estimation aims to estimate all the parameters of a causal model.

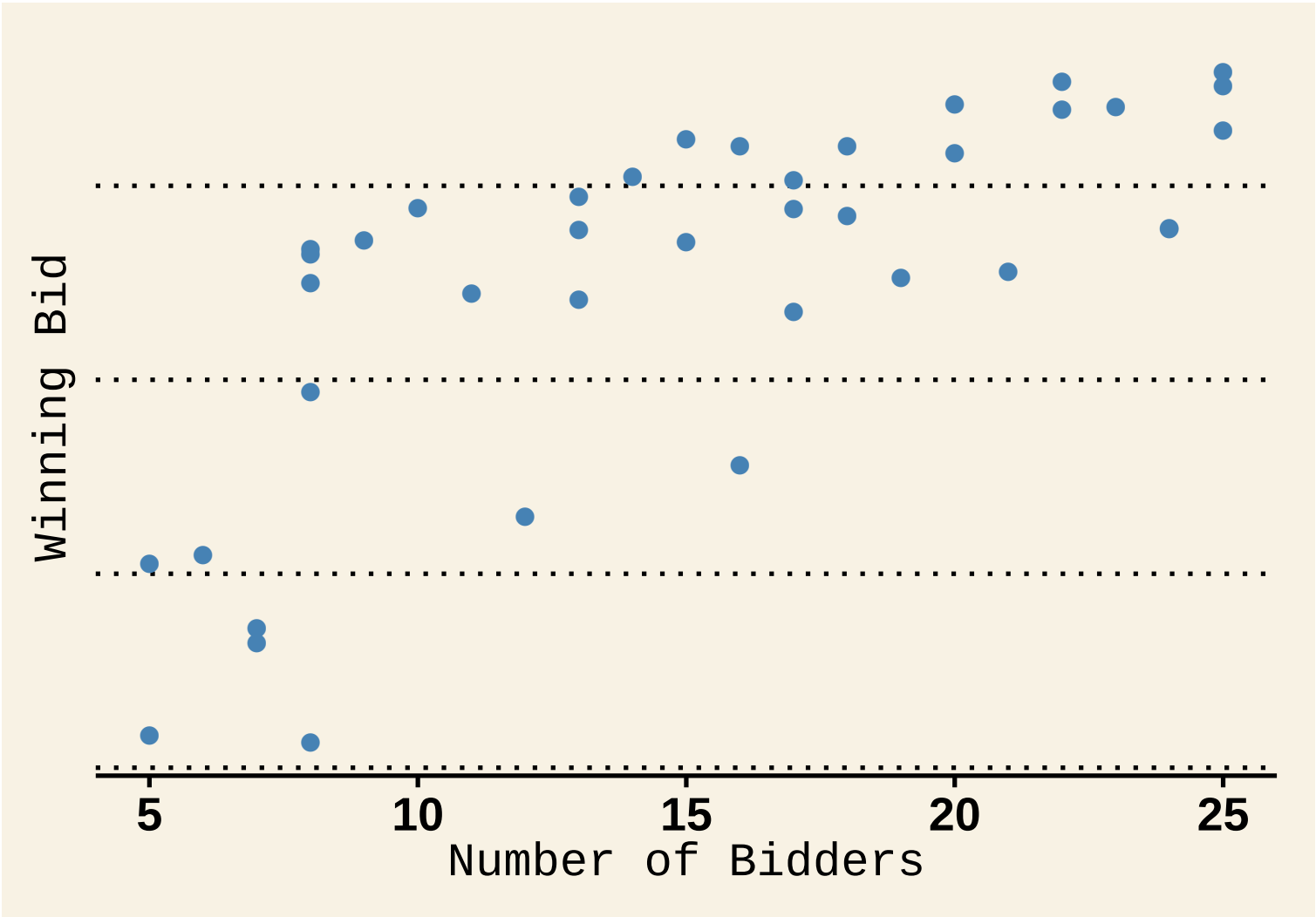
Structual Estimation

Scientific vs. Statistical Model

If you want to predict where Mars will be in the night sky[a], 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.

[a] Example 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

$\Rightarrow$ Equilibrium bidding strategy:

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

, 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[a], nonparametrically estimate $G_N(.)$ and $g_N(.)$ from observed bids $\{b_1, \ldots, b_N\}$.
2. For each bidder, calculate

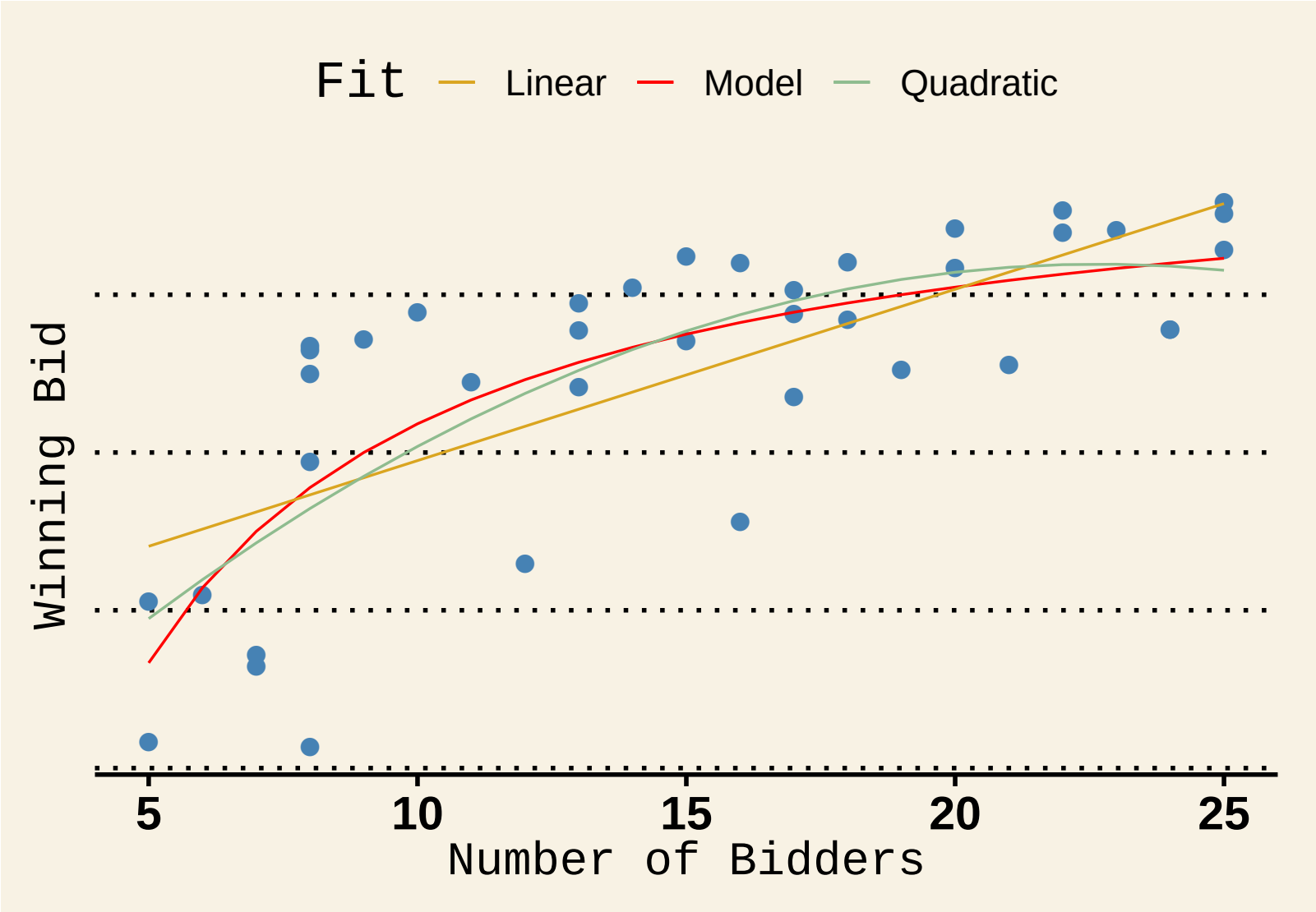
$$\widehat{v}_i = b_i + \frac{1}{N-1} \frac{\widehat{G}_N(b_i)}{\widehat{g}_N(b_i)} \tag{1}$$

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

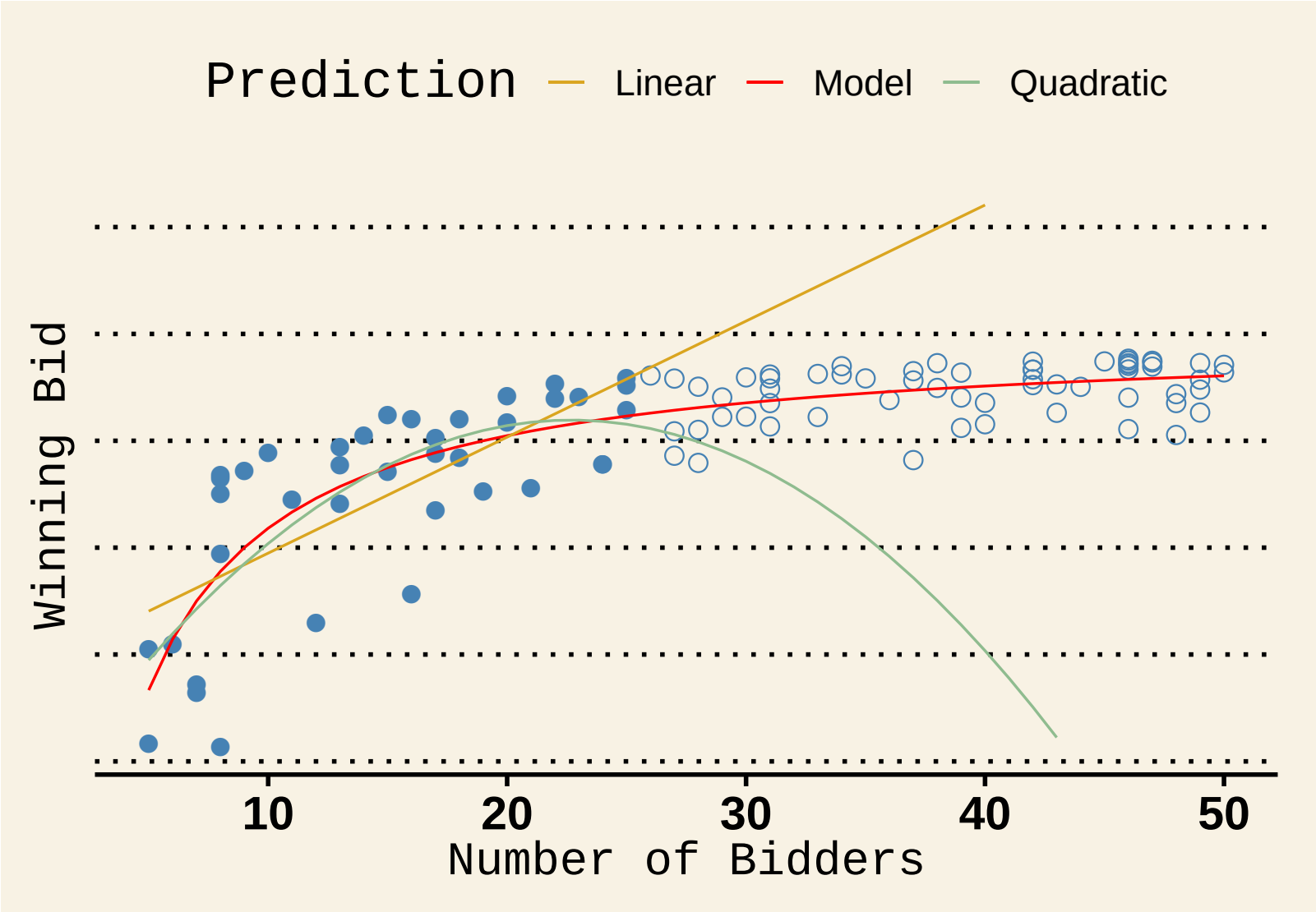
$$E[\max\{b_i\}] = E\left[\max\left\{v_i - \frac{1}{\widehat{F}(v_i)^{N-1}} \int_0^{v_i} \widehat{F}(x)^{N-1} dx\right\}\right]$$

[a] See Guerre et al. (2000).

Auction



Auction



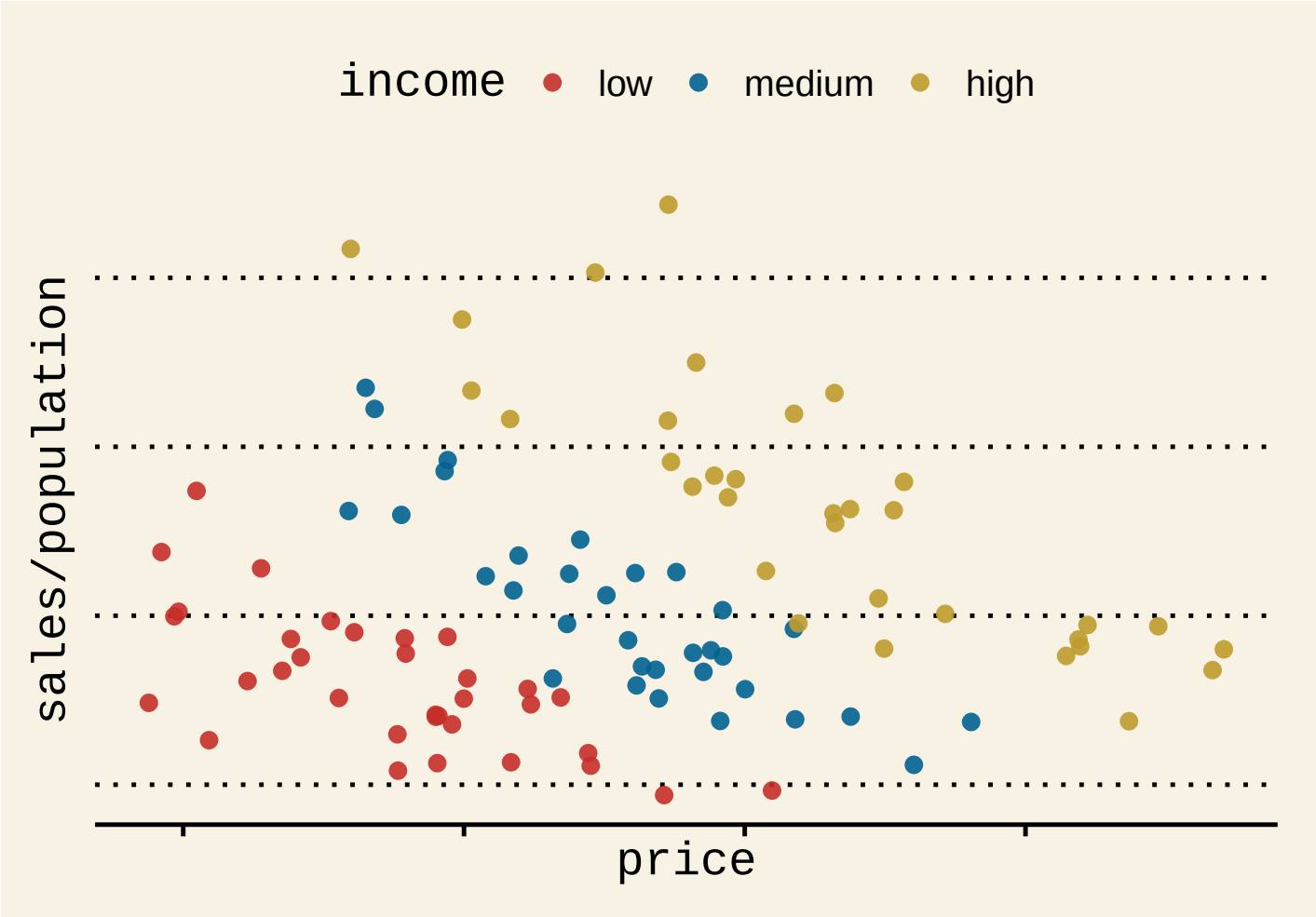
Auction

- Here, there is no confounding between N (the number of bidders) and $b_{\max}$ (the winning bid). Hence $f(N) \equiv E[b_{\max} | \operatorname{do}(N)] = E[b_{\max} | N]$.
- The learning problem is to learn $f(N)$ from data. Here, theory helps specify the functional form of $f(N)$ and therefore serves essentially 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:

$$\begin{aligned} U_{i0}^m &= \epsilon_{i0}^m \\ U_{i1}^m &= \beta_0 + \beta_1 I_m - \beta_2 p_m + \epsilon_{i1}^m \end{aligned} \tag{2}$$

, 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:

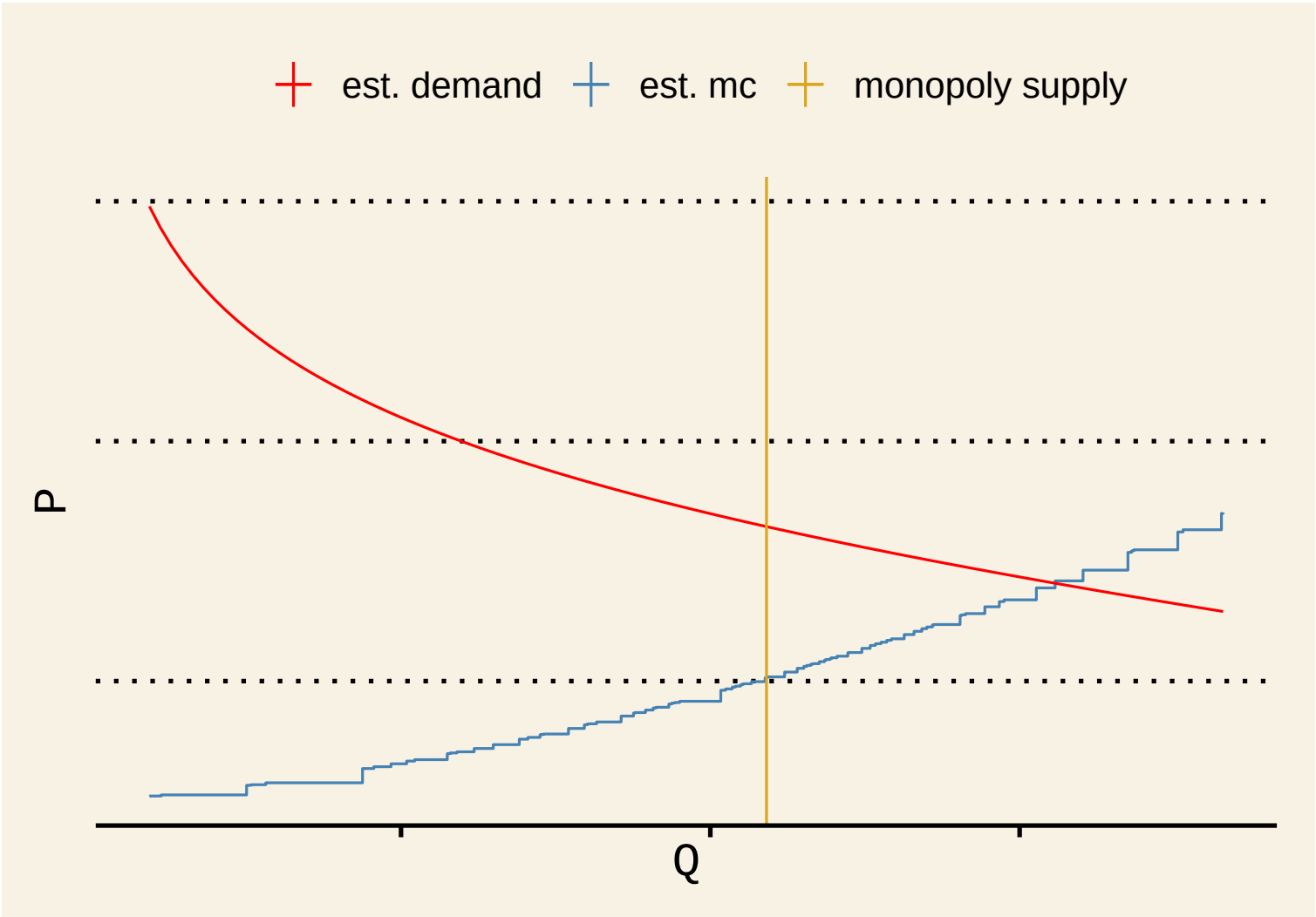
$$\max_p \{p \times q_m(p) - c(q_m(p))\} \tag{3}$$

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

(3) $\Rightarrow$

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

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 is always equal to 0, what would be the effect of $\text{do}\,(x = 1)$?
- Because structural estimation learns an entire structural model, once we have learned a model with variables $\{x_1, \ldots, x_n\}$, we can use it to generate data from the distribution $p\left(x_1, \ldots, x_n | \,\text{do}\,(x_j = a)\right)$ for any hypothetical manipulation $\text{do}\,(x_j = a)$. This is called **counterfactual simulation**.

Counterfactual Simulation

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[13], 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.

[13] Think of quantum mechanics!

Acknowledgement I

Part of this lecture is adapted from the following sources:

- Abu-Mostafa, Y. S., M. Magdon-Ismail, and H. Lin. 2012. *Learning from Data*. AMLBook.
- Angrist, J. D. and J. Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.
- Blei, D. M. *Interacting with data*. Lecture at Princeton University, retrieved on 2017.01.01. [link]
- Doré, G. *The Dore Illustrations for Dante's Divine Comedy*. Dover Publications, 1st edition (1976).
- Goodfellow, I., Y. Bengio, and A. Courville. 2016. *Deep Learning*. The MIT Press.
- Hernán, M. A. and J. M. Robins. 2019. *Causal Inference*. CRC Press.

Acknowledgement II

- 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.
- Silva, R. *Causal Inference in Machine Learning*. Talk at Imperial College London, retrieved on 2017.01.01. [link]
- Van der Schaar, M. and S. Flaxman. *Statistical Machine Learning*. Lecture at Oxford University, retrieved on 2018.01.01. [link]
- Varian, H. R., *Machine Learning and Econometrics.* Talk at Google, retrieved on 2017.01.01. [link]

Reference

- Deaton, A. and N. Cartwright. 2018. "Understanding and misunderstanding randomized controlled trials," *Social Science & Medicine*, 210.
- Guerre, E., I. Perrigne, and Q. Vuong. 2000. "Optimal Nonparametric Estimation of First-Price Auctions," *Econometrica,* 68(3).
- Russell, B., 1912. *The Problems of Philosophy*. Arc Manor, Rockville, MD (2008).
- Shalizi, C. R. 2016. *Advanced Data Analysis from an Elementary Point of View*. Manuscript.