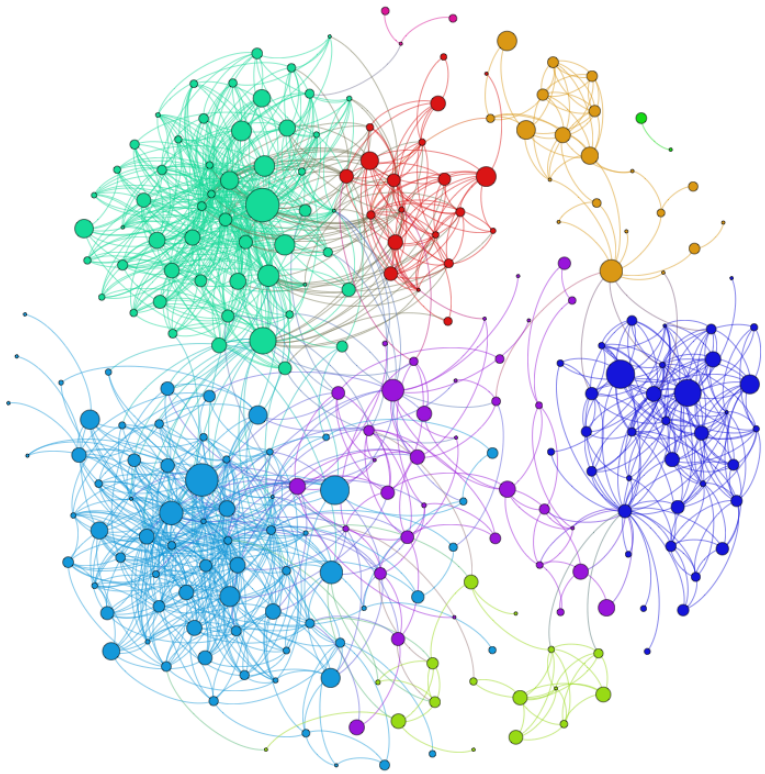
Data Analysis for Economics

ä Modern Introduction

Jiaming Mao

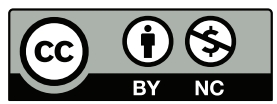
Xiamen University



This version: Spring 2022

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**	

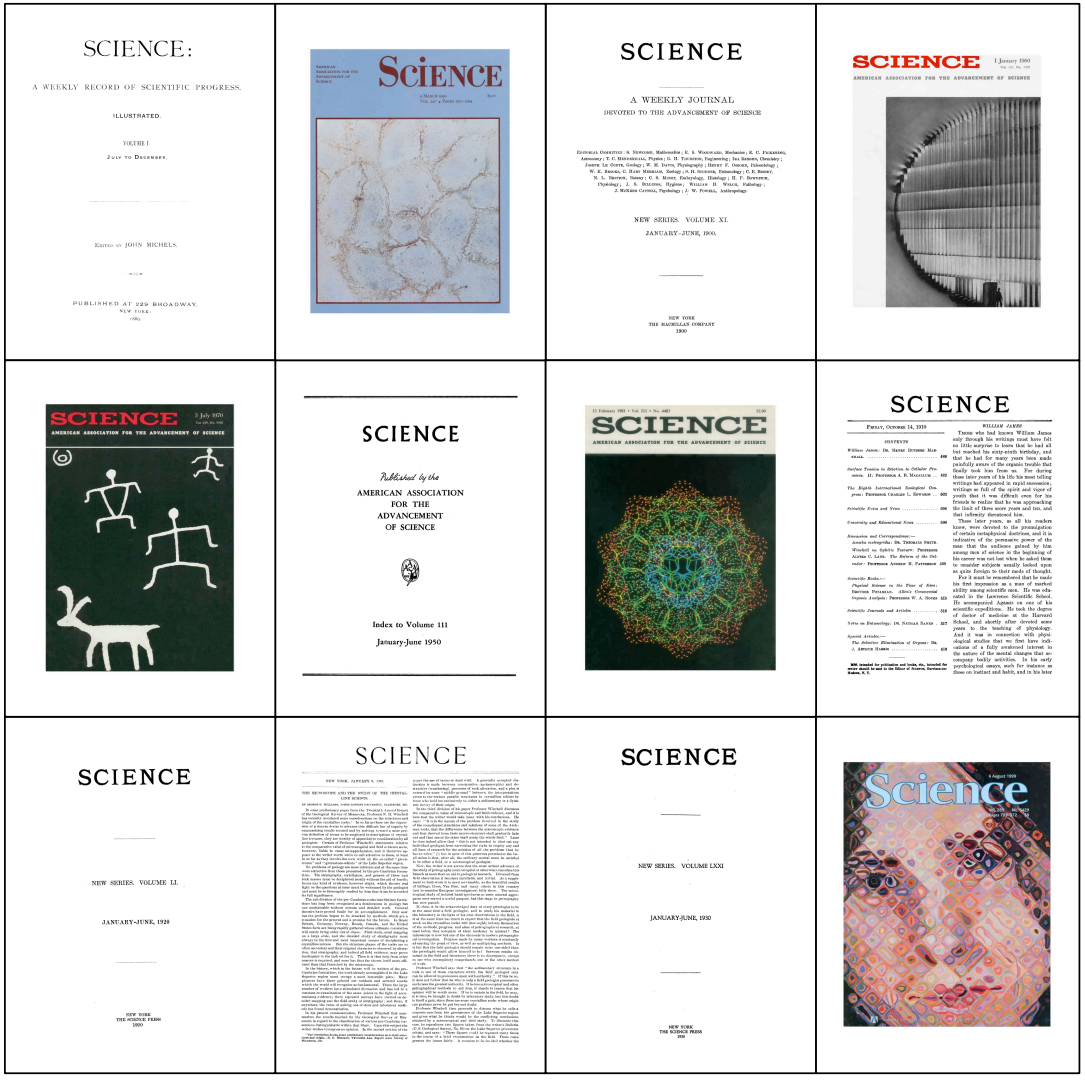
Purchase histories

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	

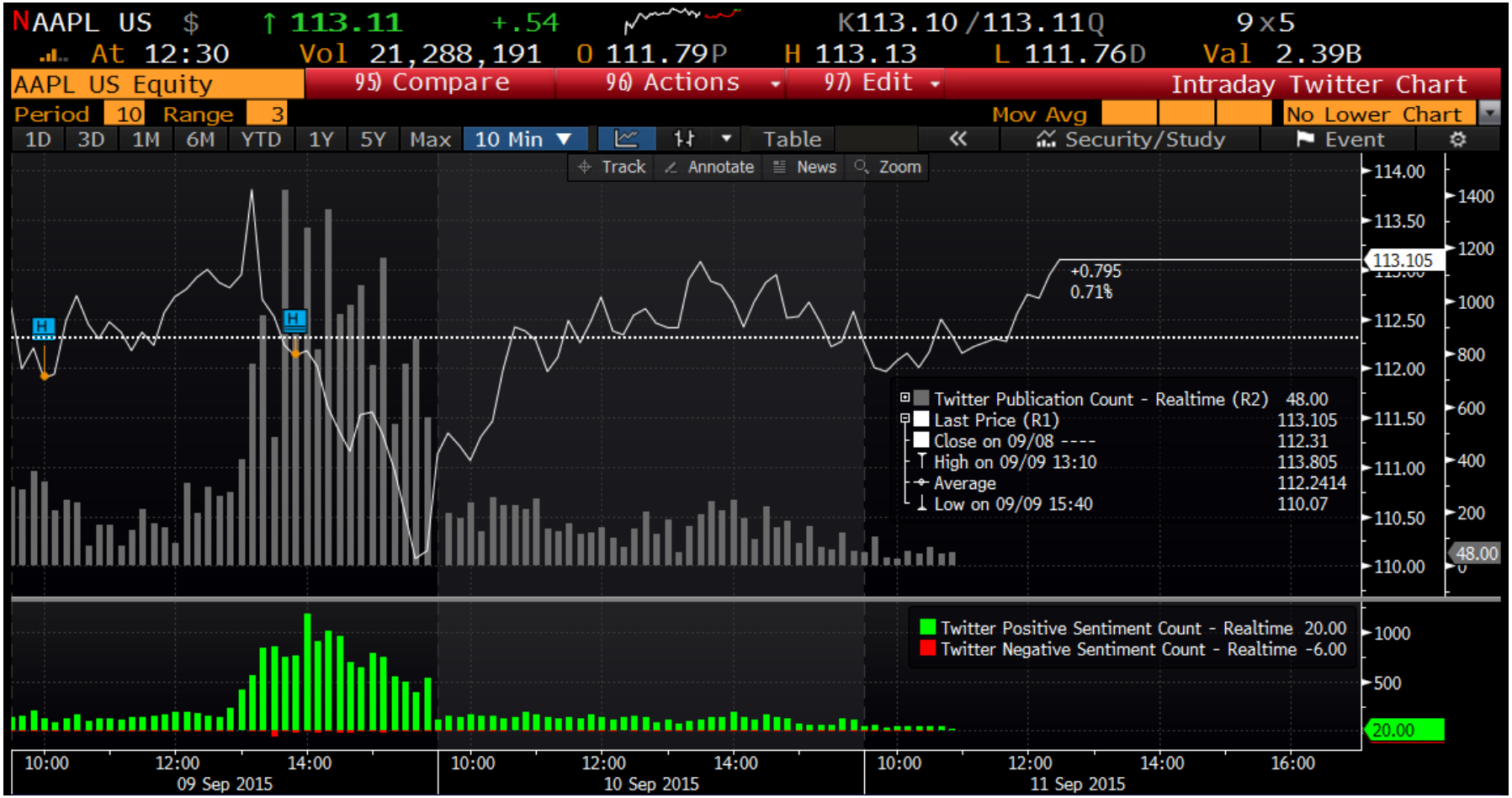
User ratings

Data are everywhere



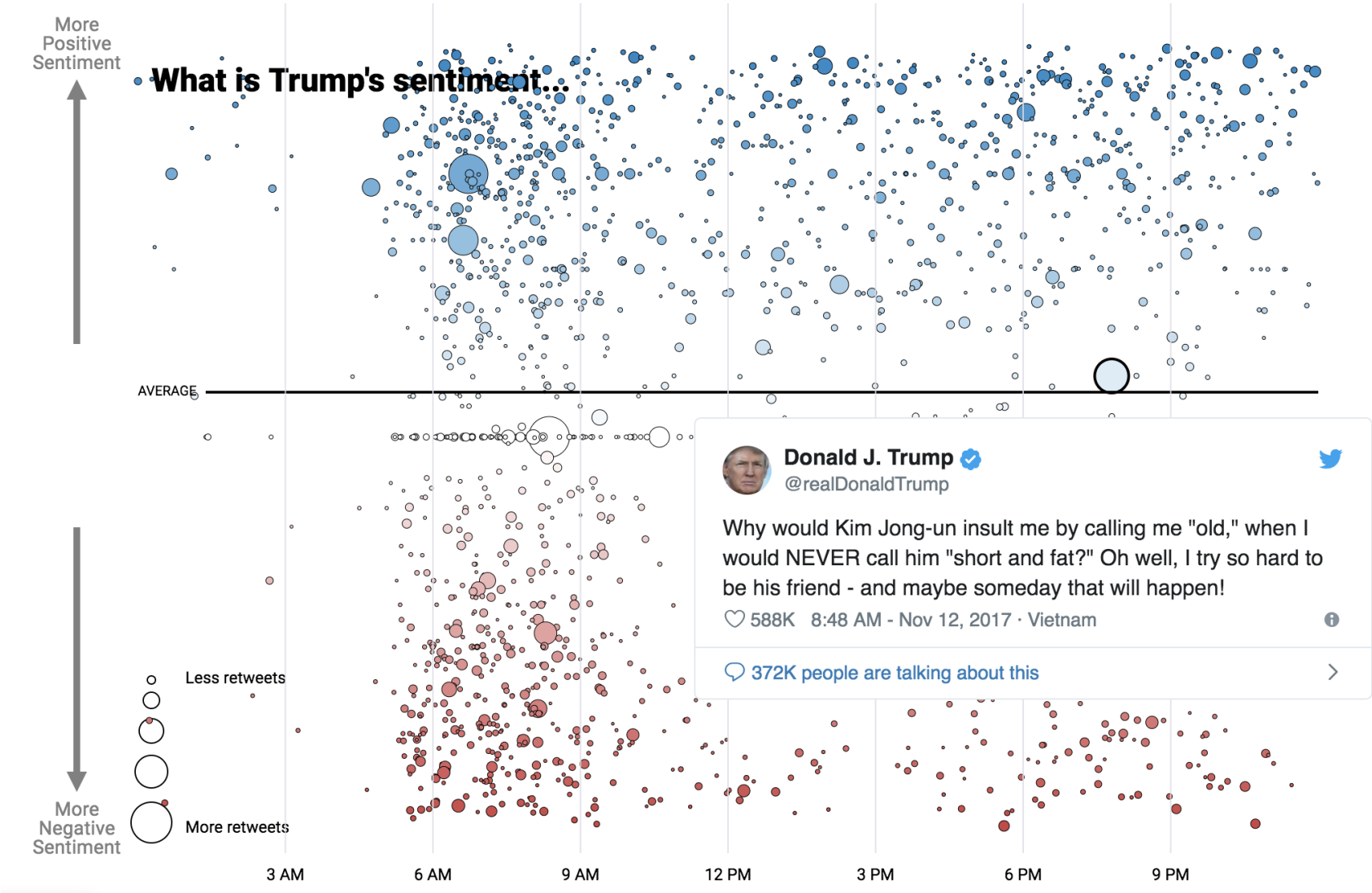
Document collections

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.

Pattern Discovery

Classification









Which one is a chair?

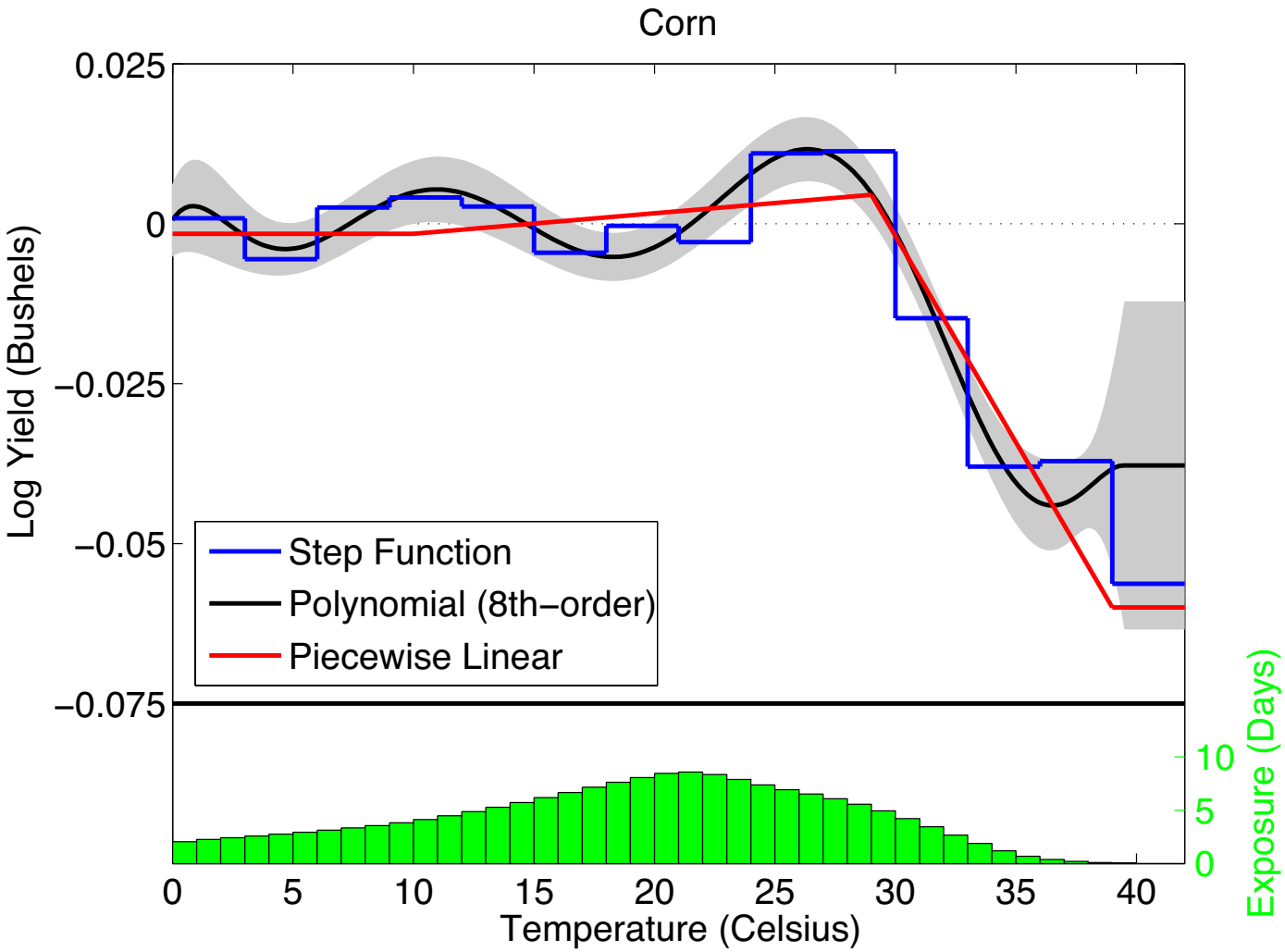
Pattern Discovery

Classification

- Which product will a consumer buy?
- Which market will a firm enter?
- Which political candidate will an individual vote for?

Pattern Discovery

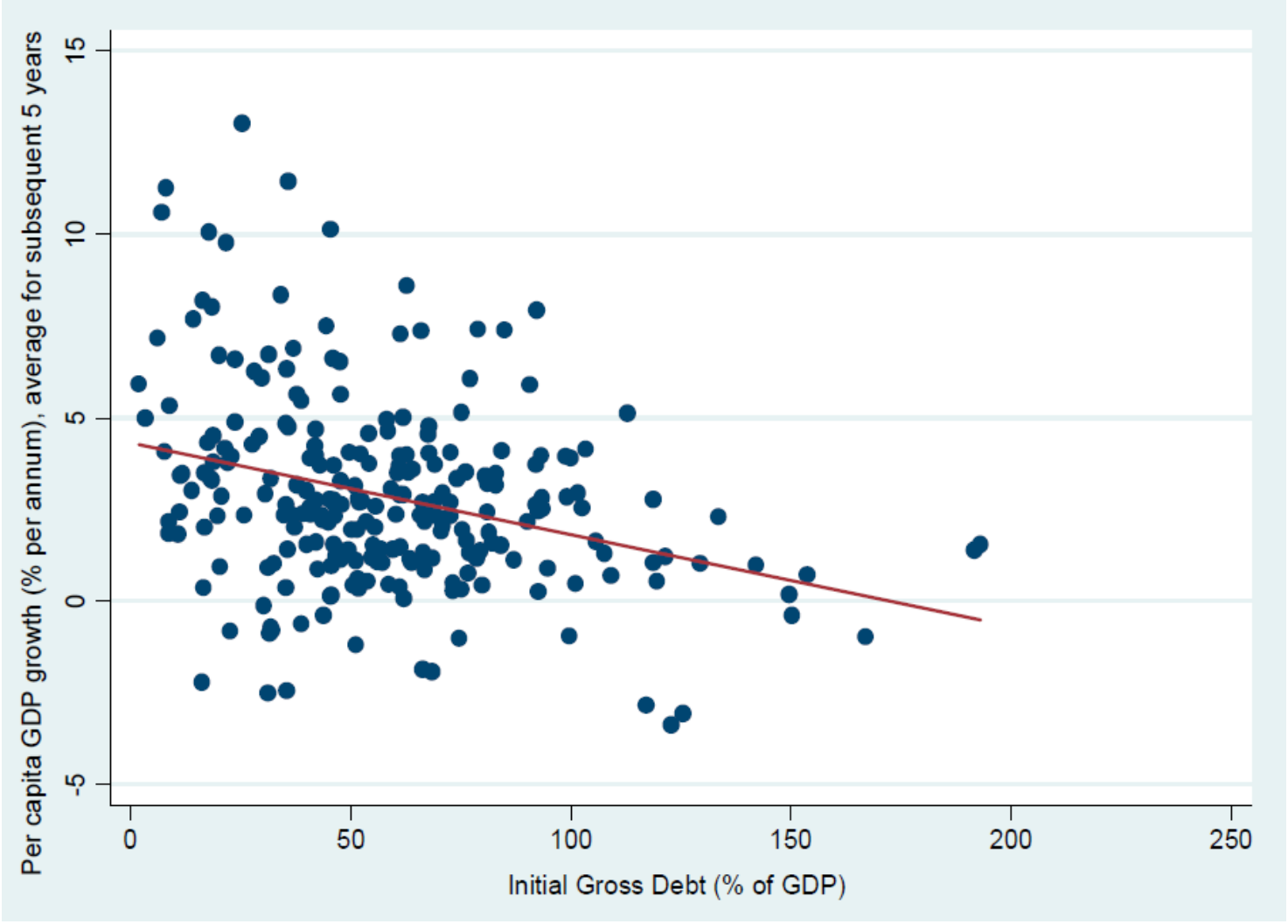
Regression



Temperature and crop yield

Pattern Discovery

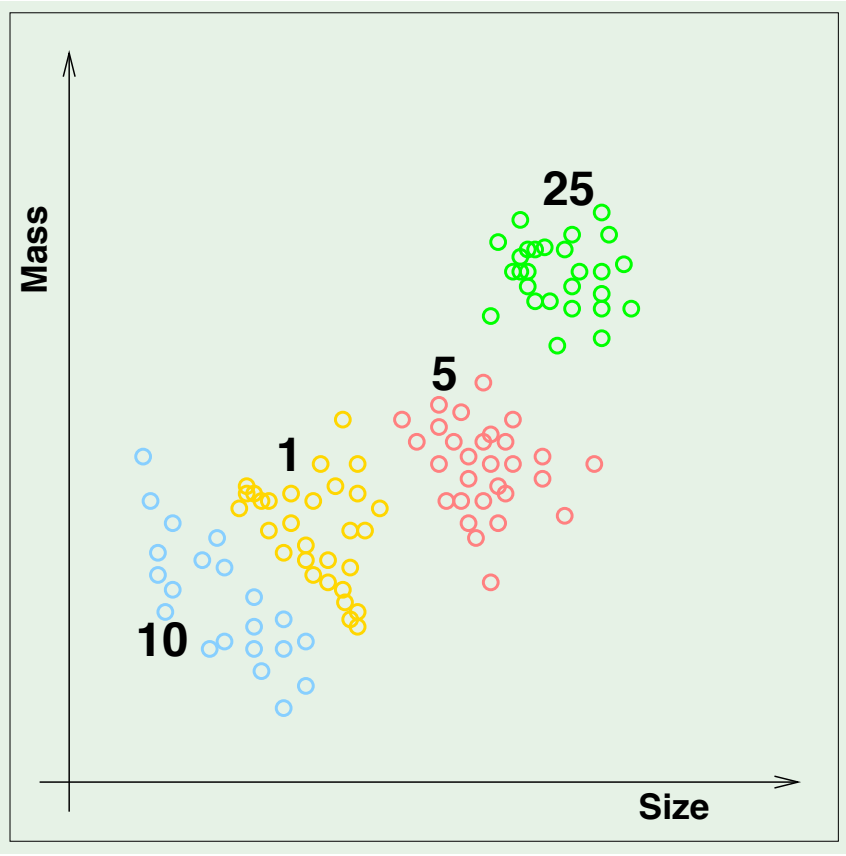
Regression

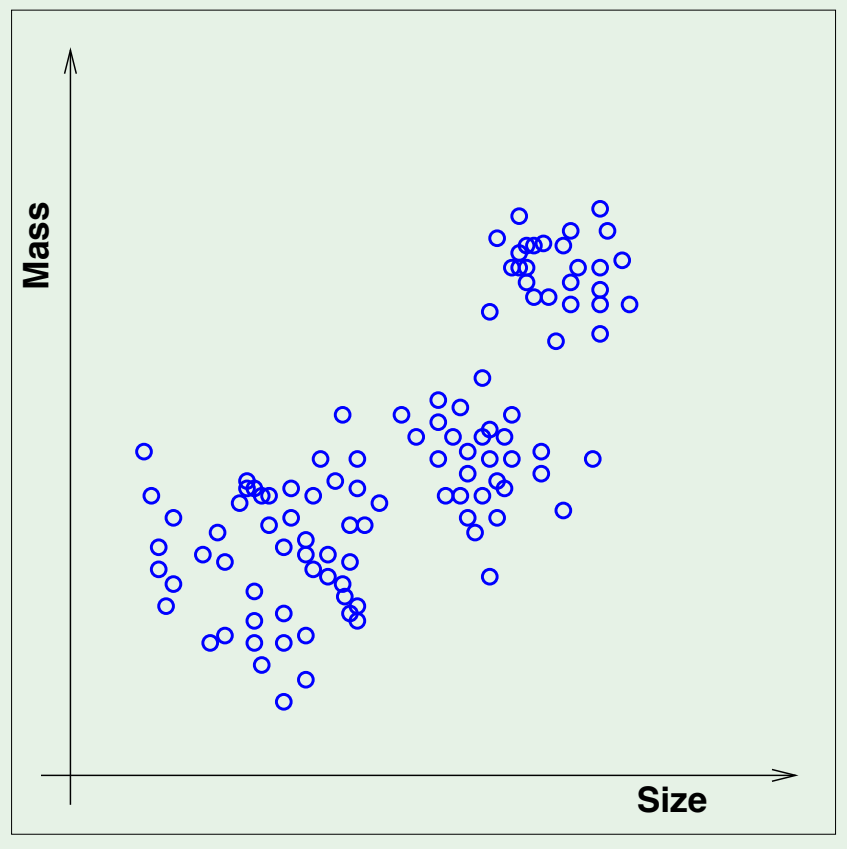


Government debt and GDP growth

Pattern Discovery

Unsupervised Learning



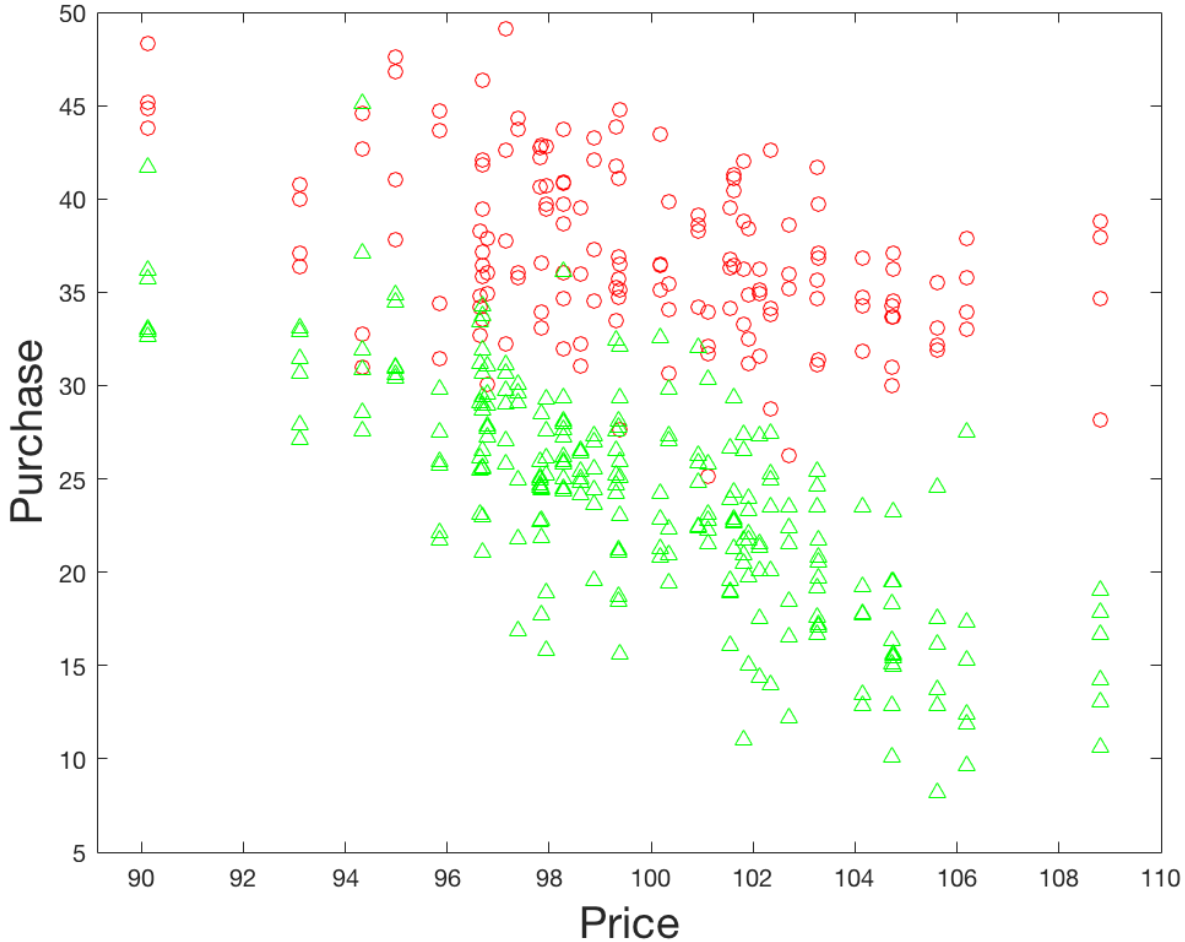


Vending machine coin recognition

Left: supervised learning; Right: unsupervised learning

Pattern Discovery

Unsupervised Learning



Consumer demand

Machine Learning Methods for Economic Applications

"In this paper, we review and apply several popular methods from the machine learning literature to the problem of demand estimation ... we compare these methods to standard econometric models that are used by practitioners to study demand ... we used sales data on salty snacks [from] scanner panel data from grocery stores ... In our results, we find that the six models we use from the statistics and computer science literature predict demand out of sample in standard metrics much more accurately than a panel data or logistic model." – Bajari et al. (2015)

Machine Learning Methods for Economic Applications

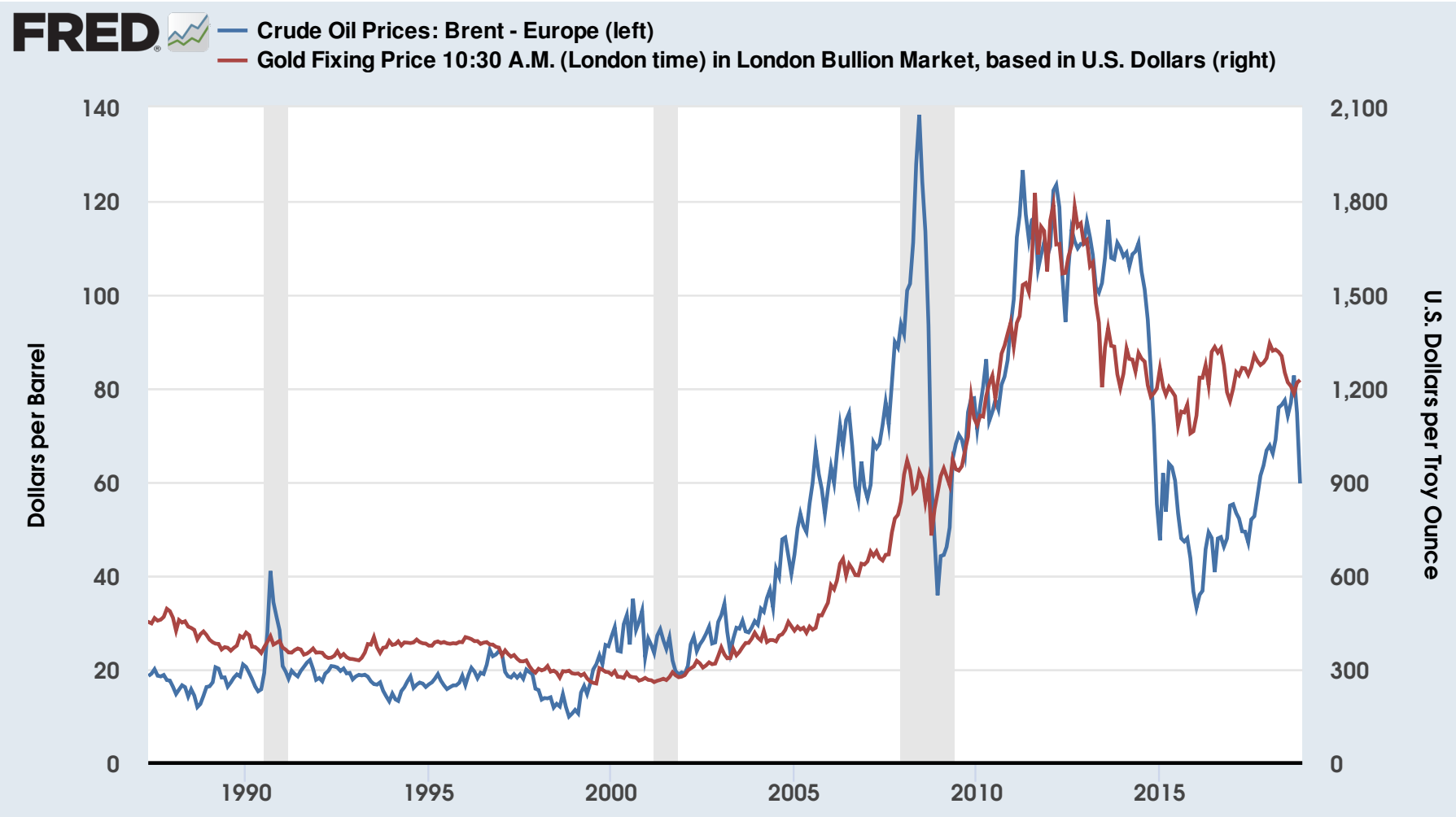
	Validation		Out-of-Sample		
	RMSE	Std. Err.	RMSE	Std. Err.	Weight
Linear	1.169	0.022	1.193	0.020	6.62%
Stepwise	0.983	0.012	1.004	0.011	12.13%
Forward Stagewise	0.988	0.013	1.003	0.012	0.00%
Lasso	1.178	0.017	1.222	0.012	0.00%
Random Forest	0.943	0.017	0.965	0.015	65.56%
SVM	1.046	0.024	1.068	0.018	15.69%
Bagging	1.355	0.030	1.321	0.025	0.00%
Logit	1.190	0.020	1.234	0.018	0.00%
Combined	0.924		0.946		100.00%
# of Obs	226,952		376,980		
Total Obs	1,510,563				
% of Total	15.0%		25.0%		

Bajari et al. (2015)

Causal Inference

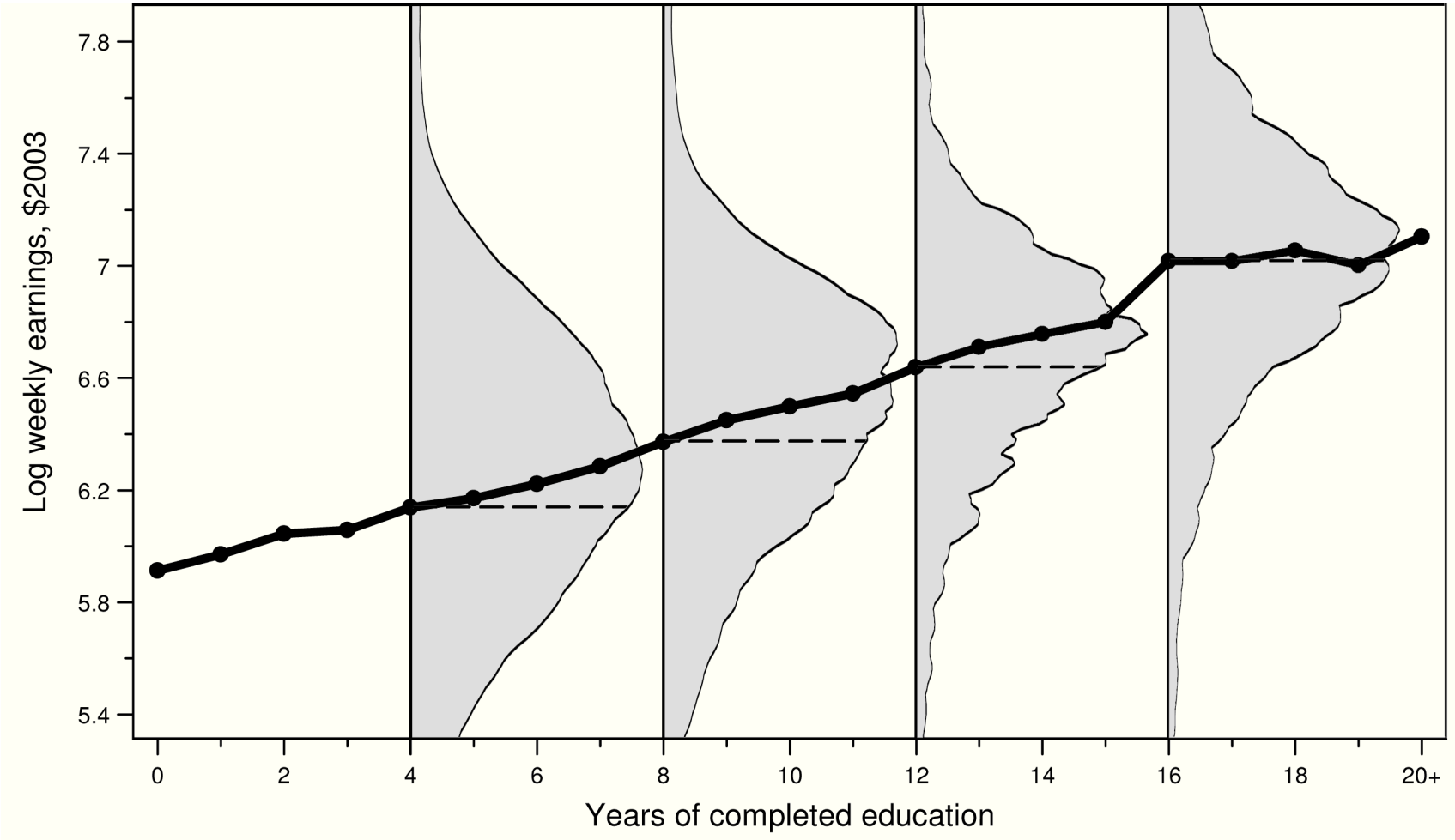
Learning patterns in the data is not enough – we want understanding.

Causal Inference



Do gold and oil prices cause each other to move or are their comovements caused by something else?

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 income taxes on labor supply
- Effect of zoning regulations on housing prices
- Effect of environmental regulations on pollution emission
- ...

Artificial Intelligence

- Research on causal inference methodologies has taken on new importance with the development of artificial intelligence (AI).
- 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.

Artificial Intelligence

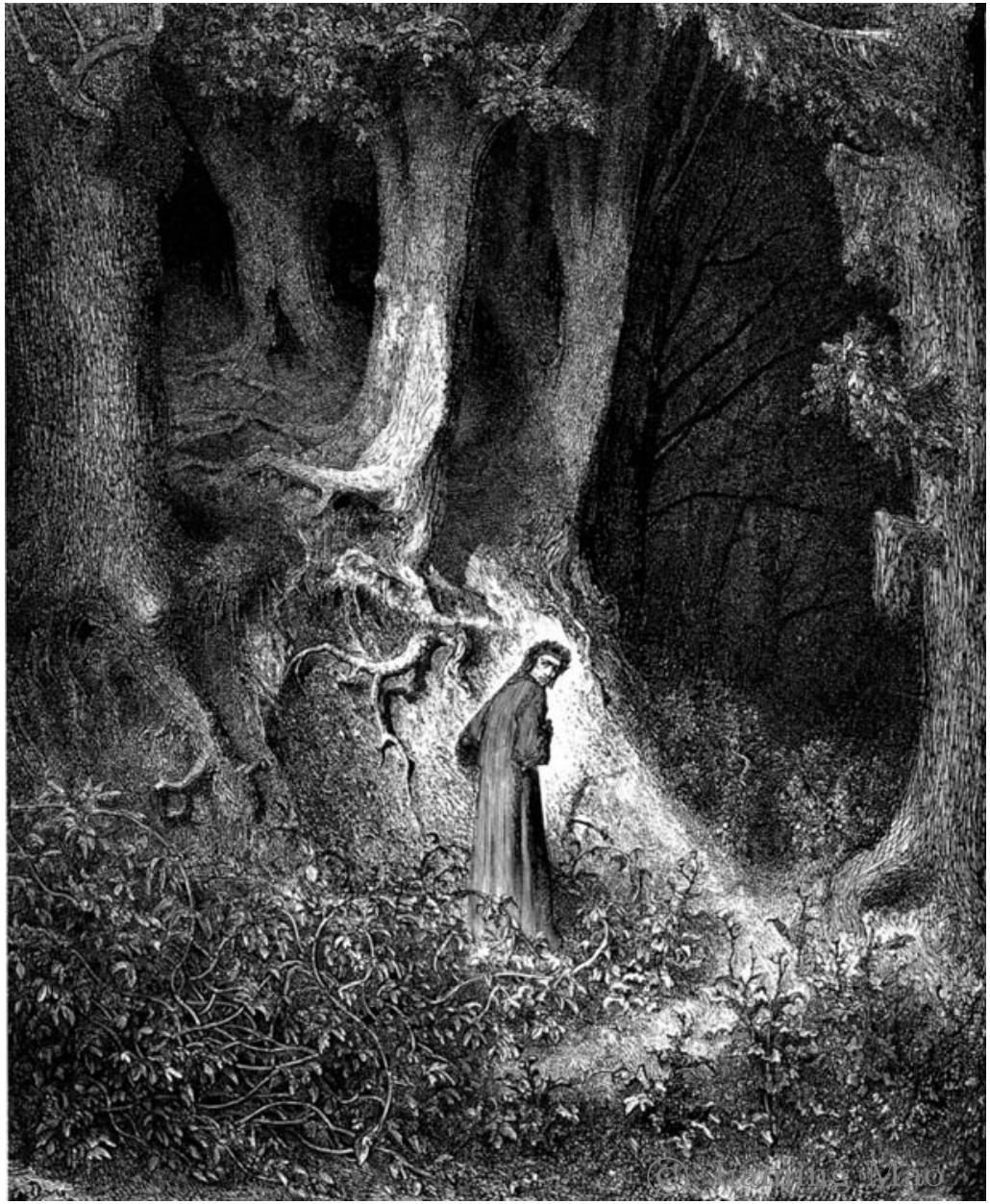
Road Map

1. Statistical Learning

2. Causal Inference

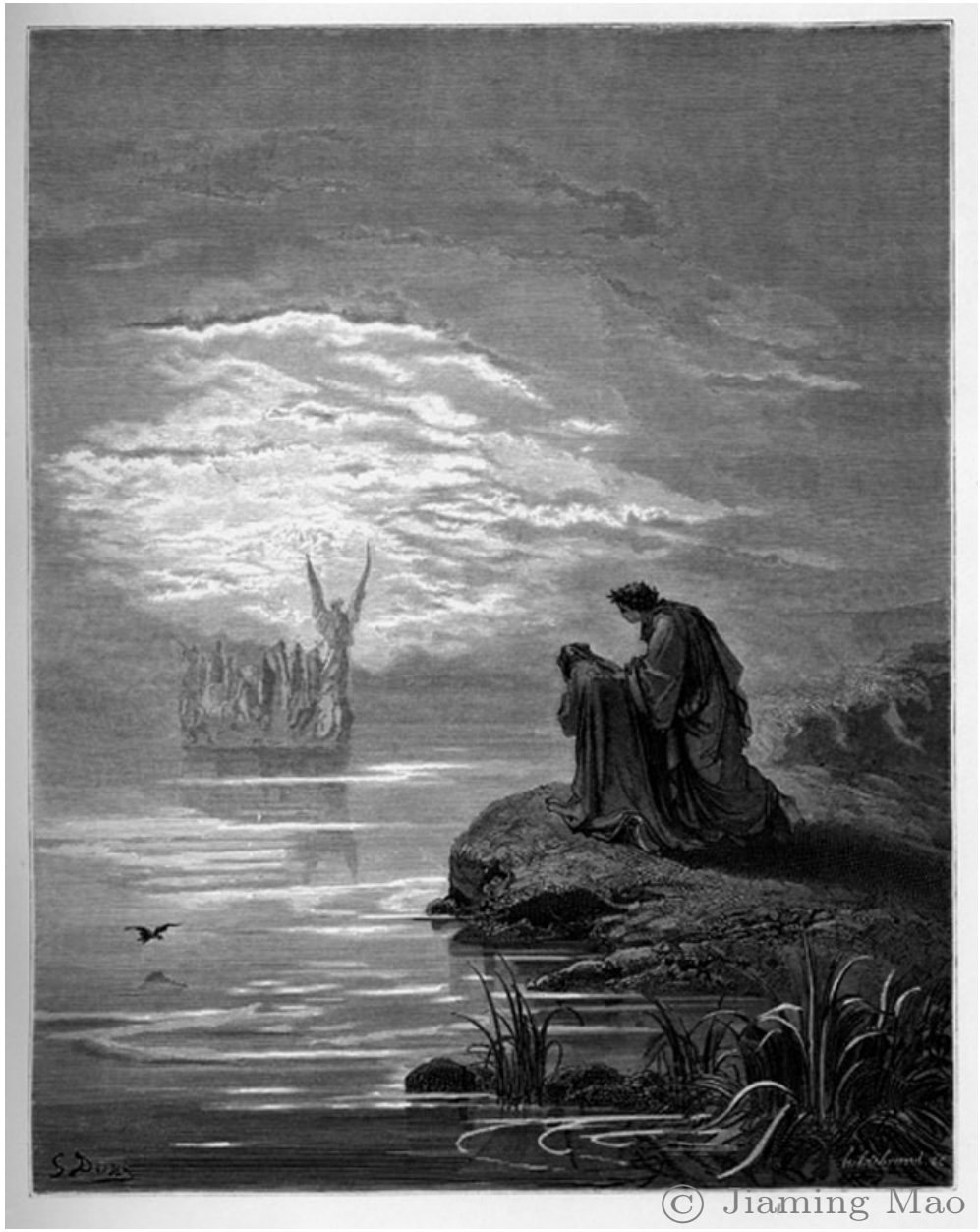
Road Map

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



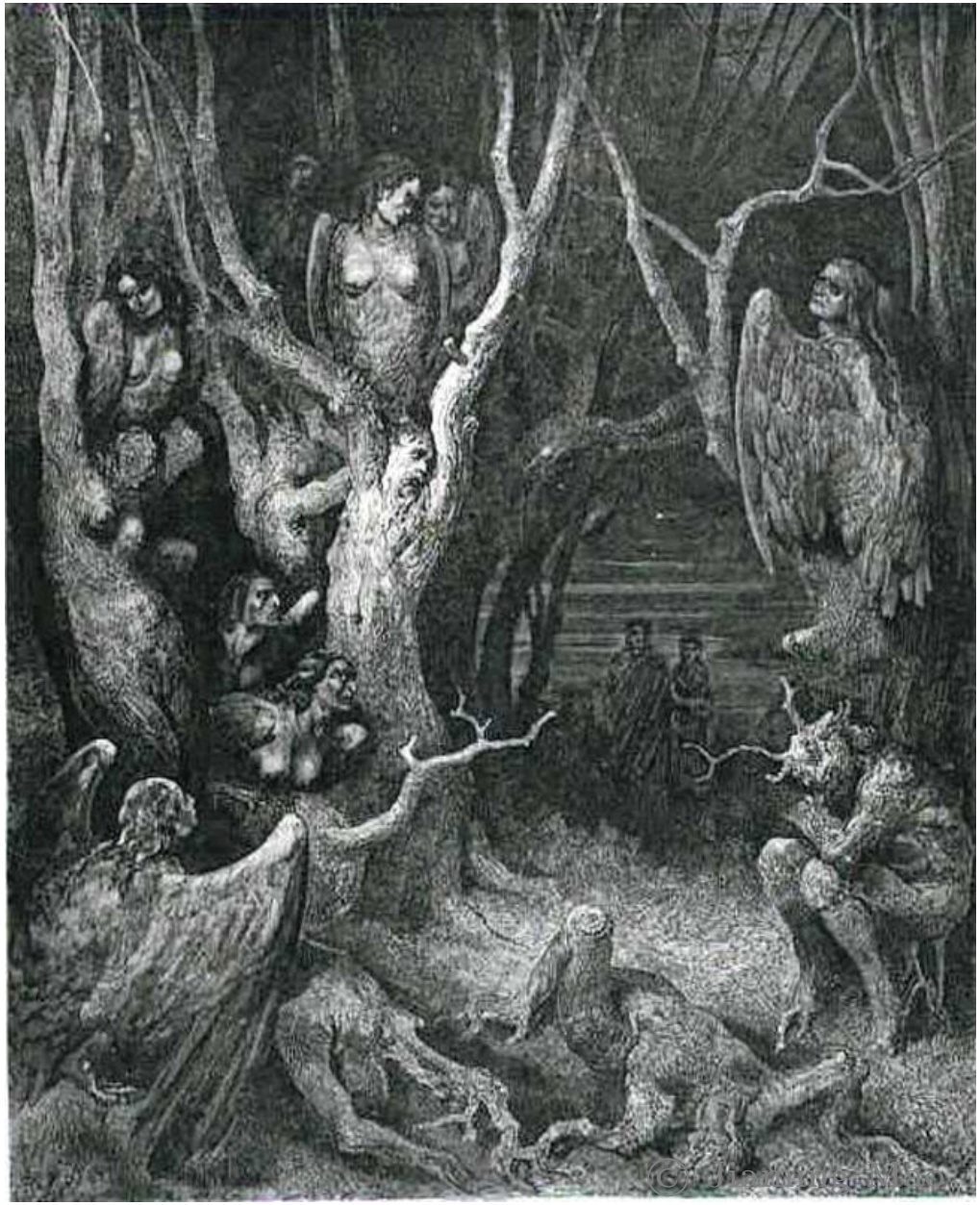
Road Map

... 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.

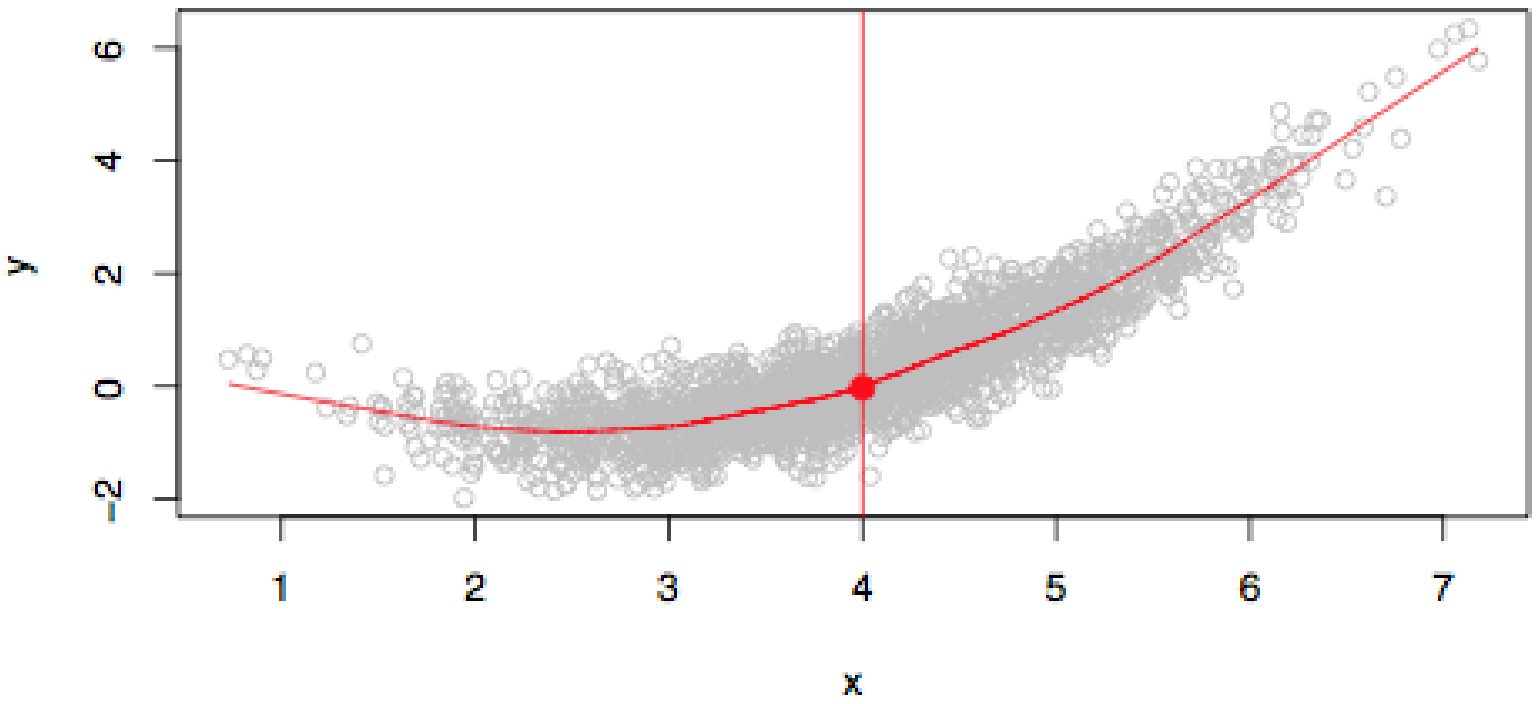
What is the function f that produces the best prediction of y given x?

- Depends on how we measure "best." Common choice: minimizing the expected squared-error loss[2] $\mathbb{E}\left[(y - f(x))^2\right] \Rightarrow f(x) = \mathbb{E}[y|x]$.
- $f(x) = \mathbb{E}[y|x]$ 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.

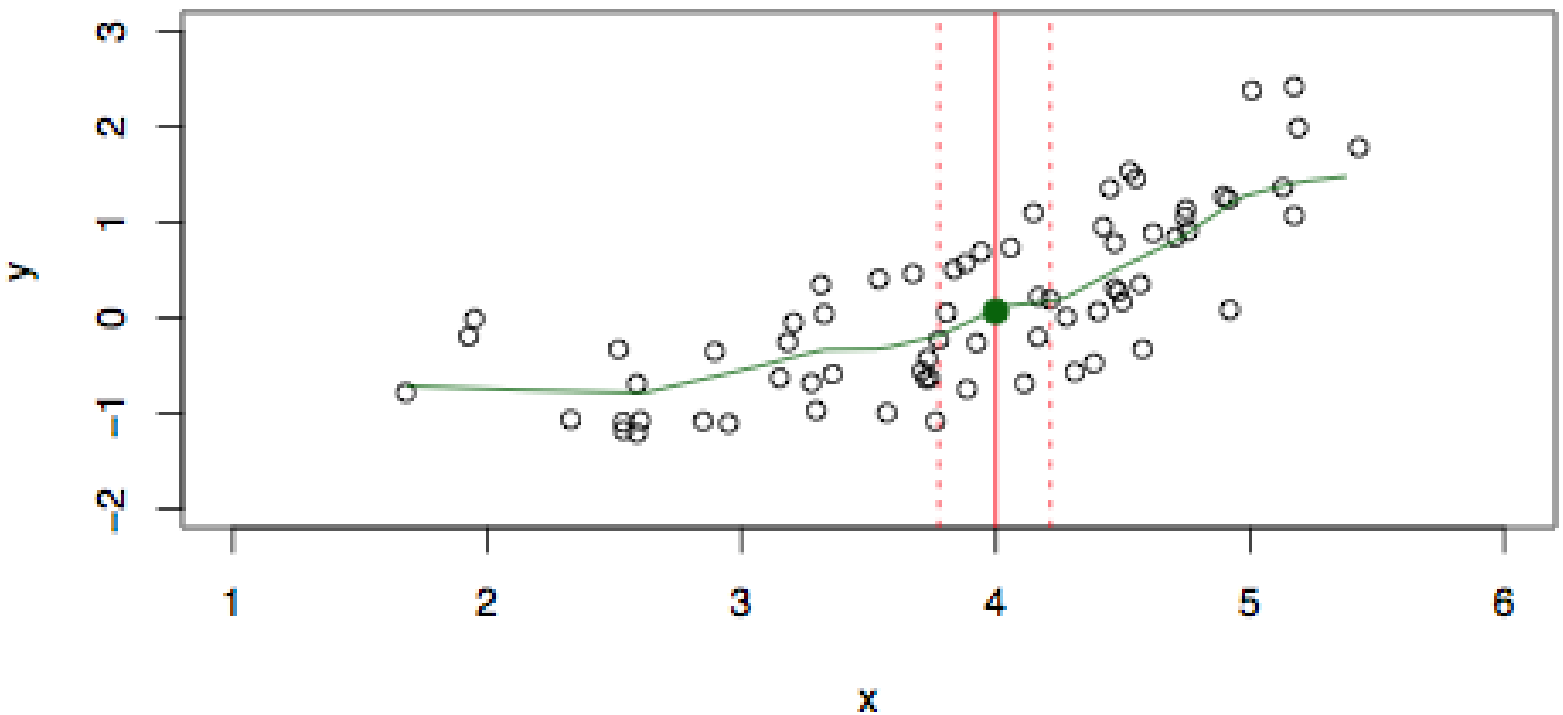
Learning f



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

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.



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

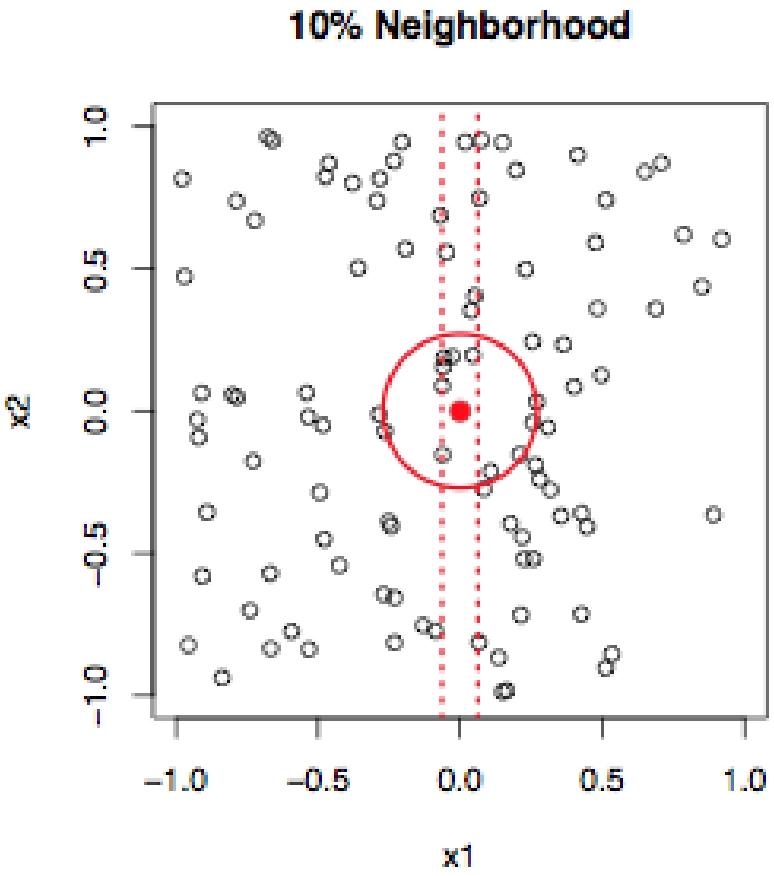
, where $\mathcal{N}\,(x)$ is some neighborhood of x.

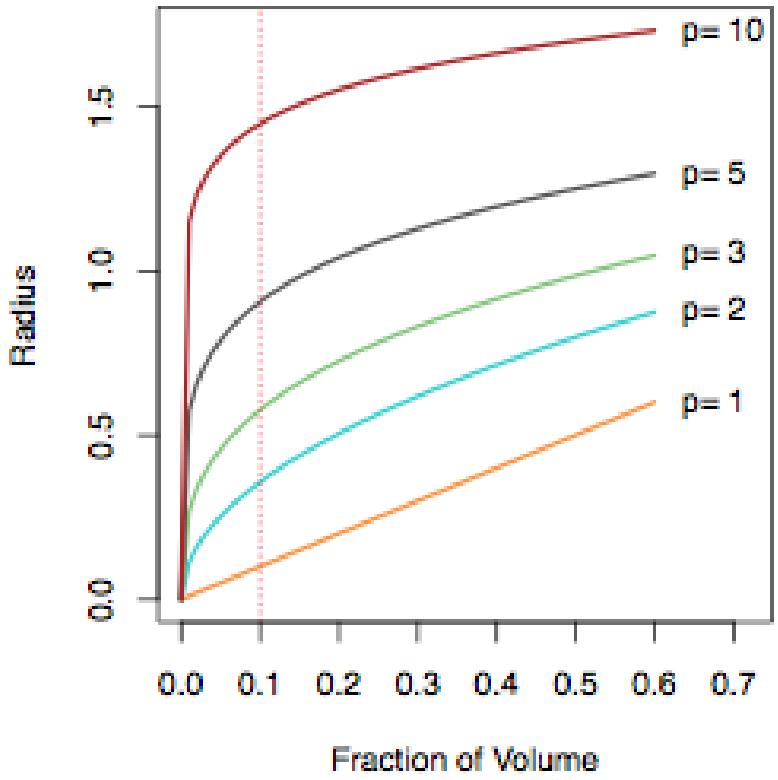
Learning f

- 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

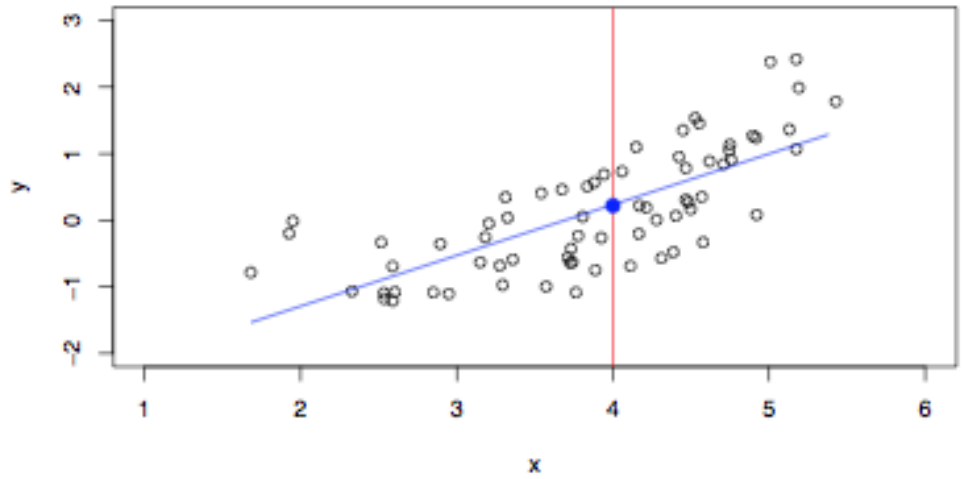
Learning f

- **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.

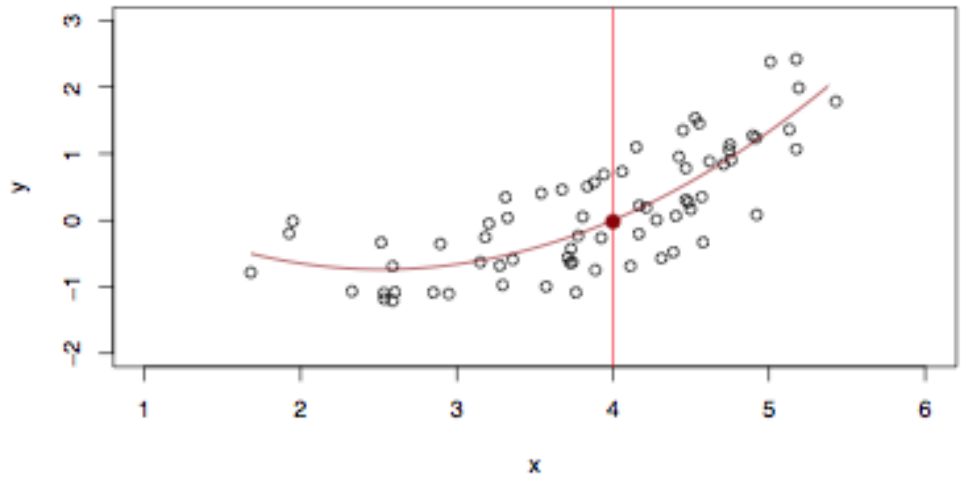
[5] We will use the terms "statistical method" and "statistical model" interchangeably.

[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.

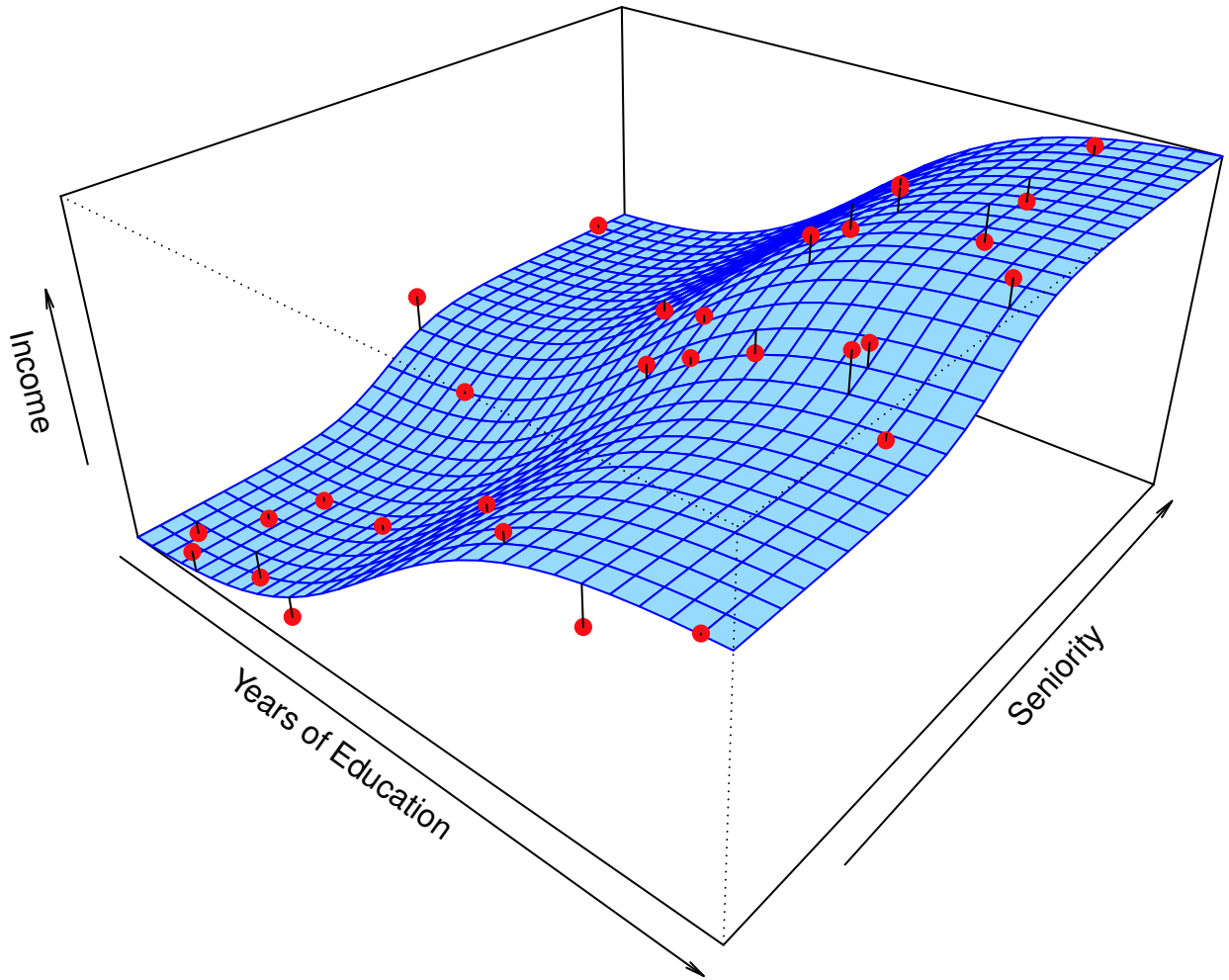
A linear model $\hat{f}_L(X) = \hat{\beta}_0 + \hat{\beta}_1 X$ gives a reasonable fit here



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

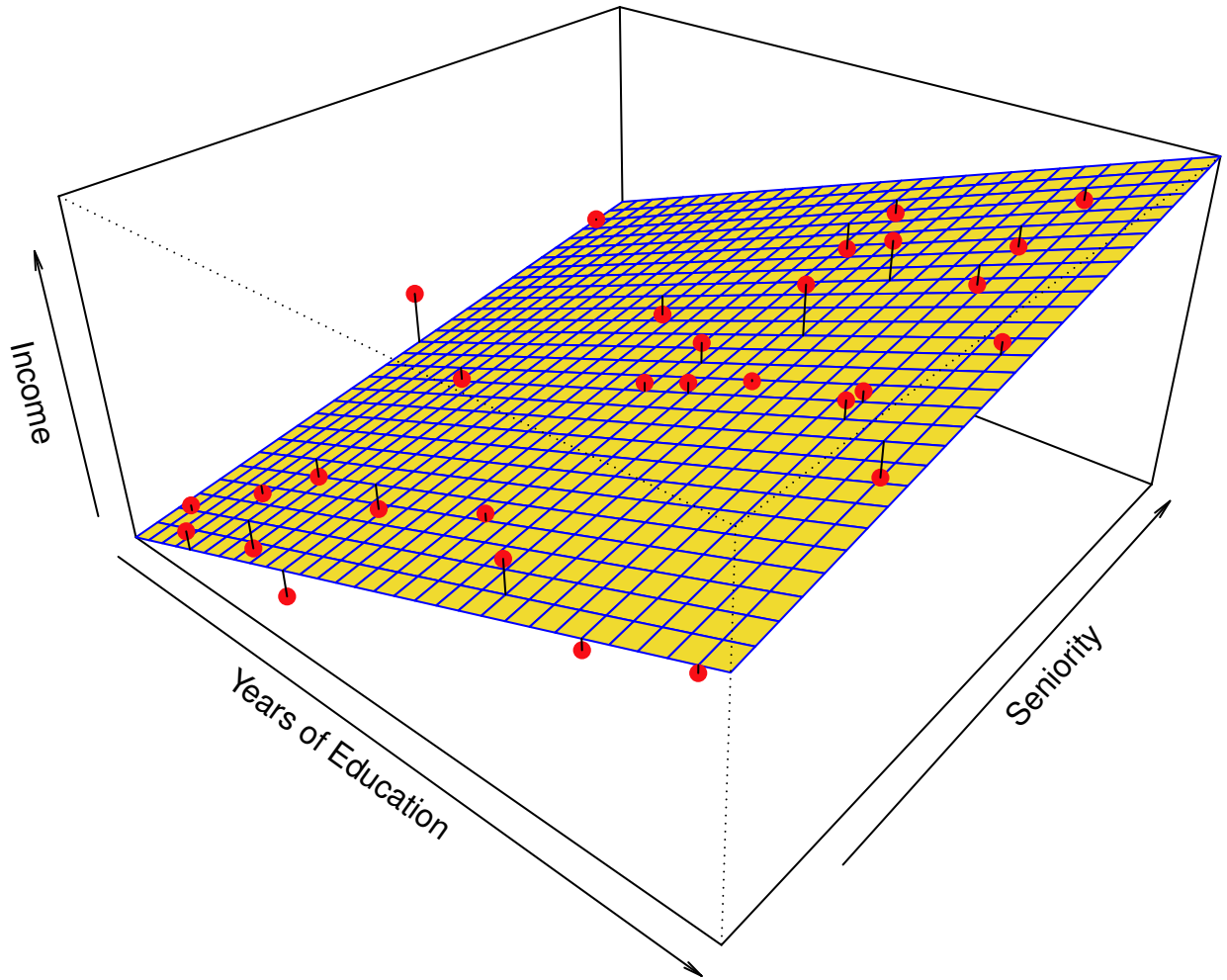


Learning f



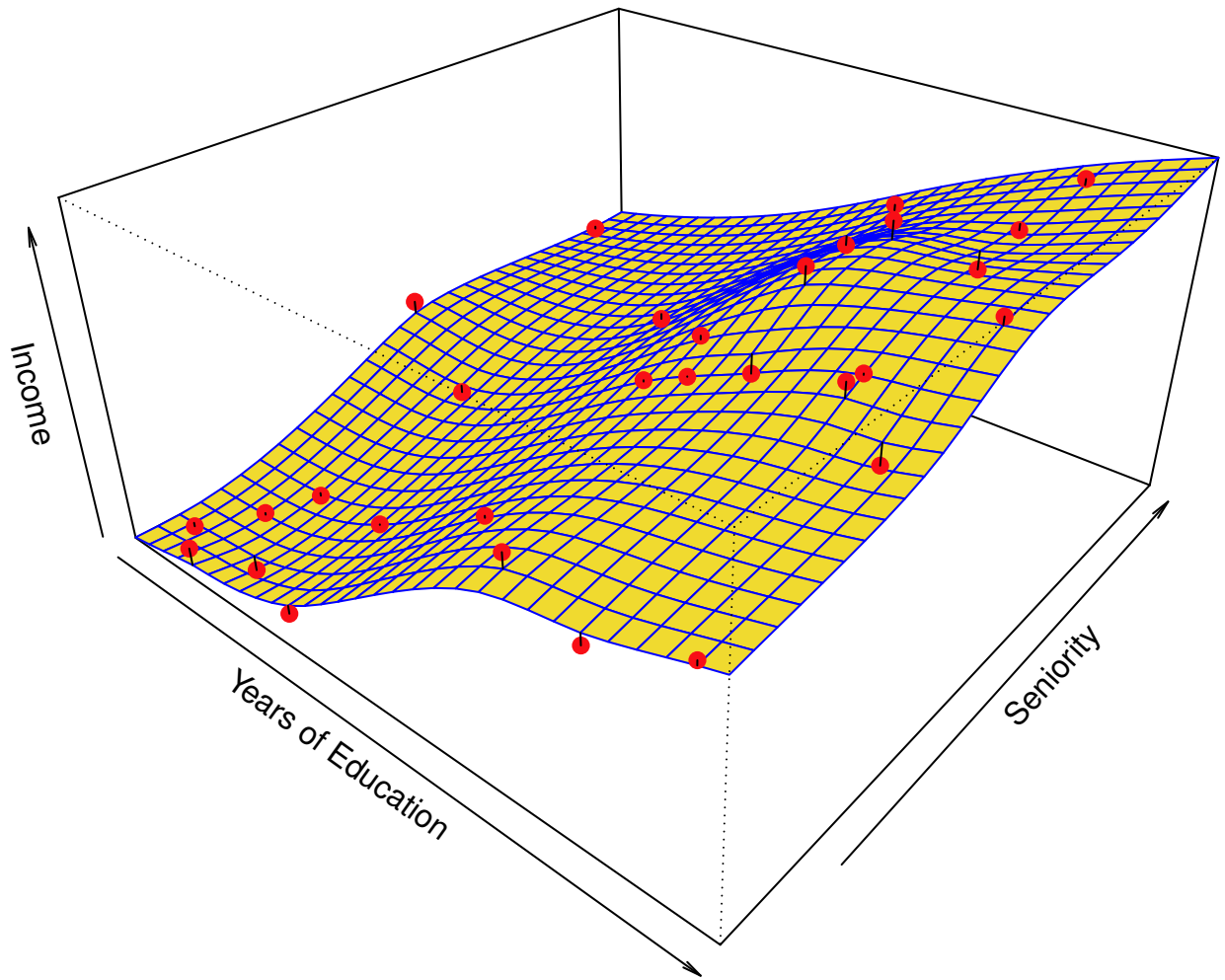
True f

Learning f



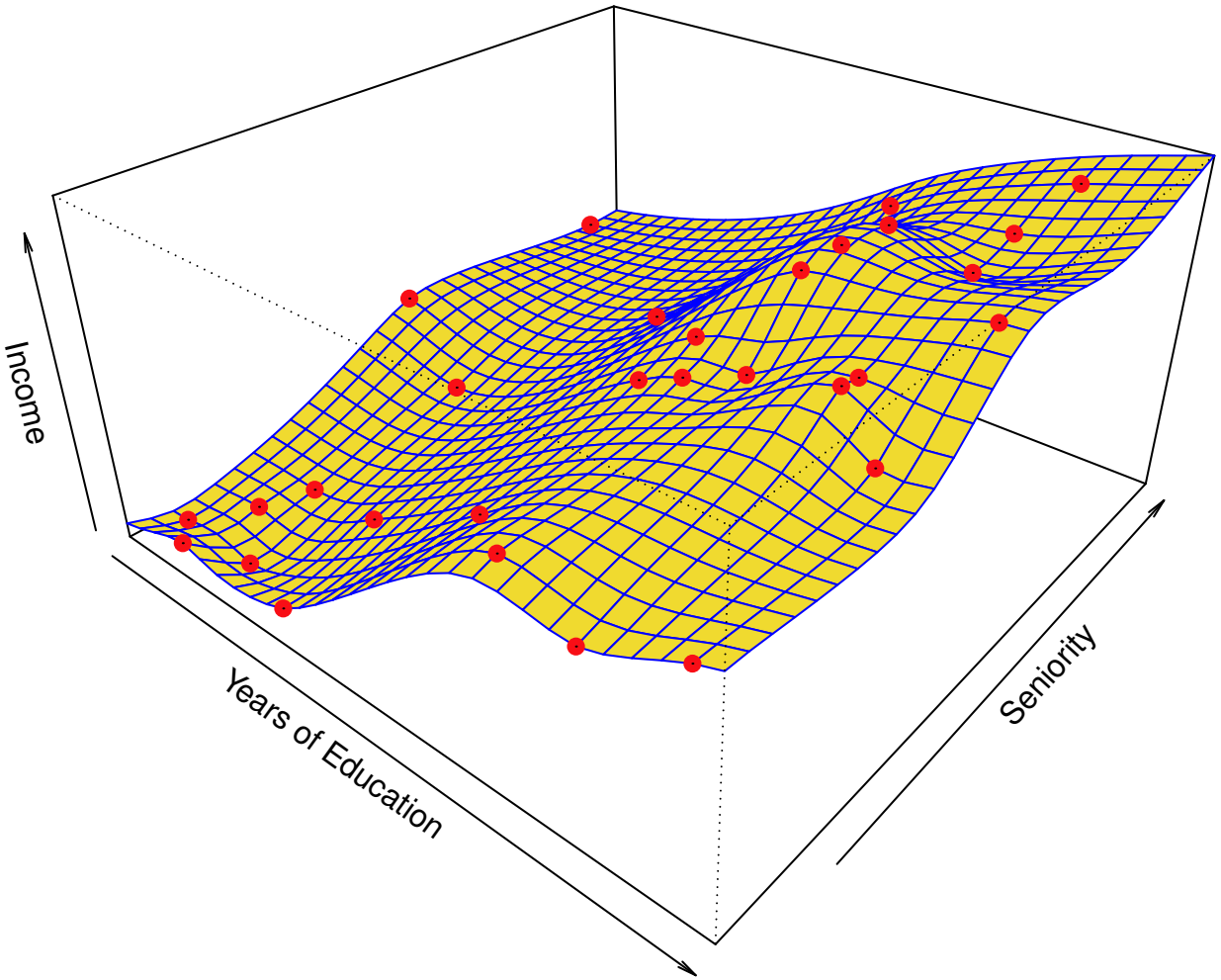
Linear Fit

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

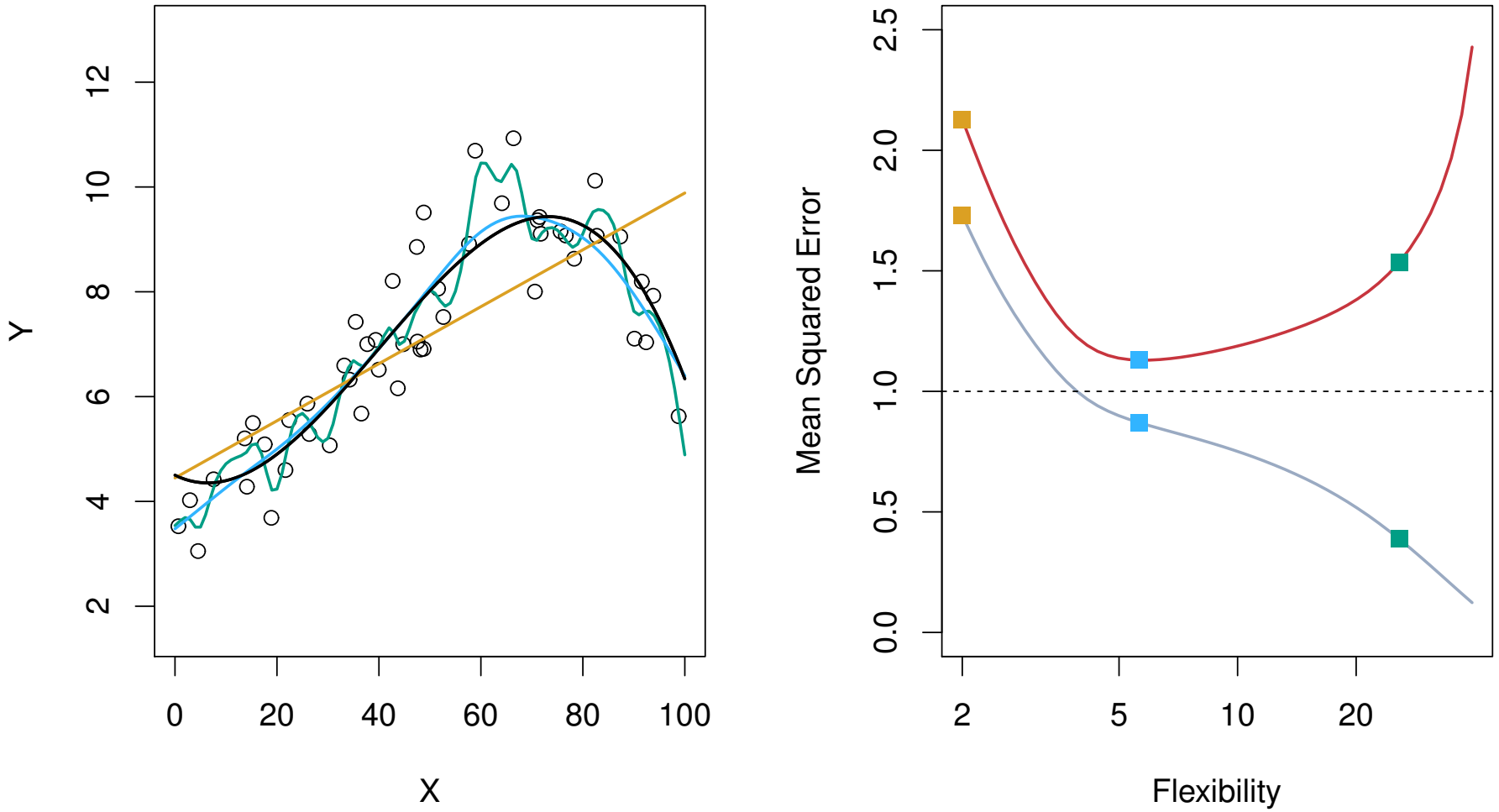
To this end, we can apply $\widehat{f}$ to a set of **test data**, $\mathcal{D}_{TE} = \{(x_1, y_1), \ldots, (x_M, y_M)\}$, and calculate the **test error**:

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

When $M \to \infty$, $\text{error}_{TE} \to \mathbb{E}\left[\left(y - \widehat{f}(x) \right)^2 \right]$[7].

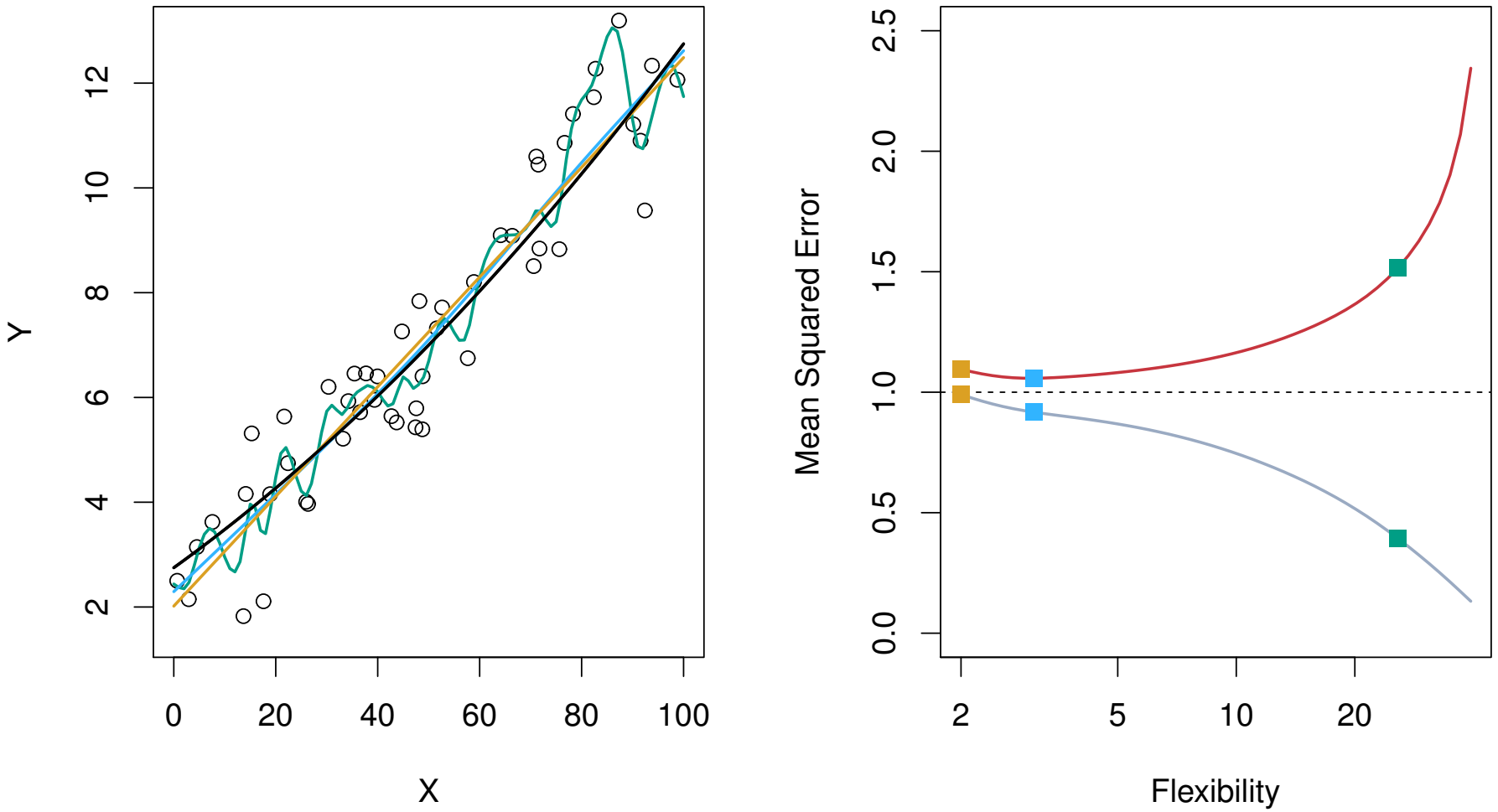
[7] $\mathbb{E}[(y - \widehat{f}(x))^2]$ is called **expected error** or **prediction error**.

Assessing the Goodness of Fit



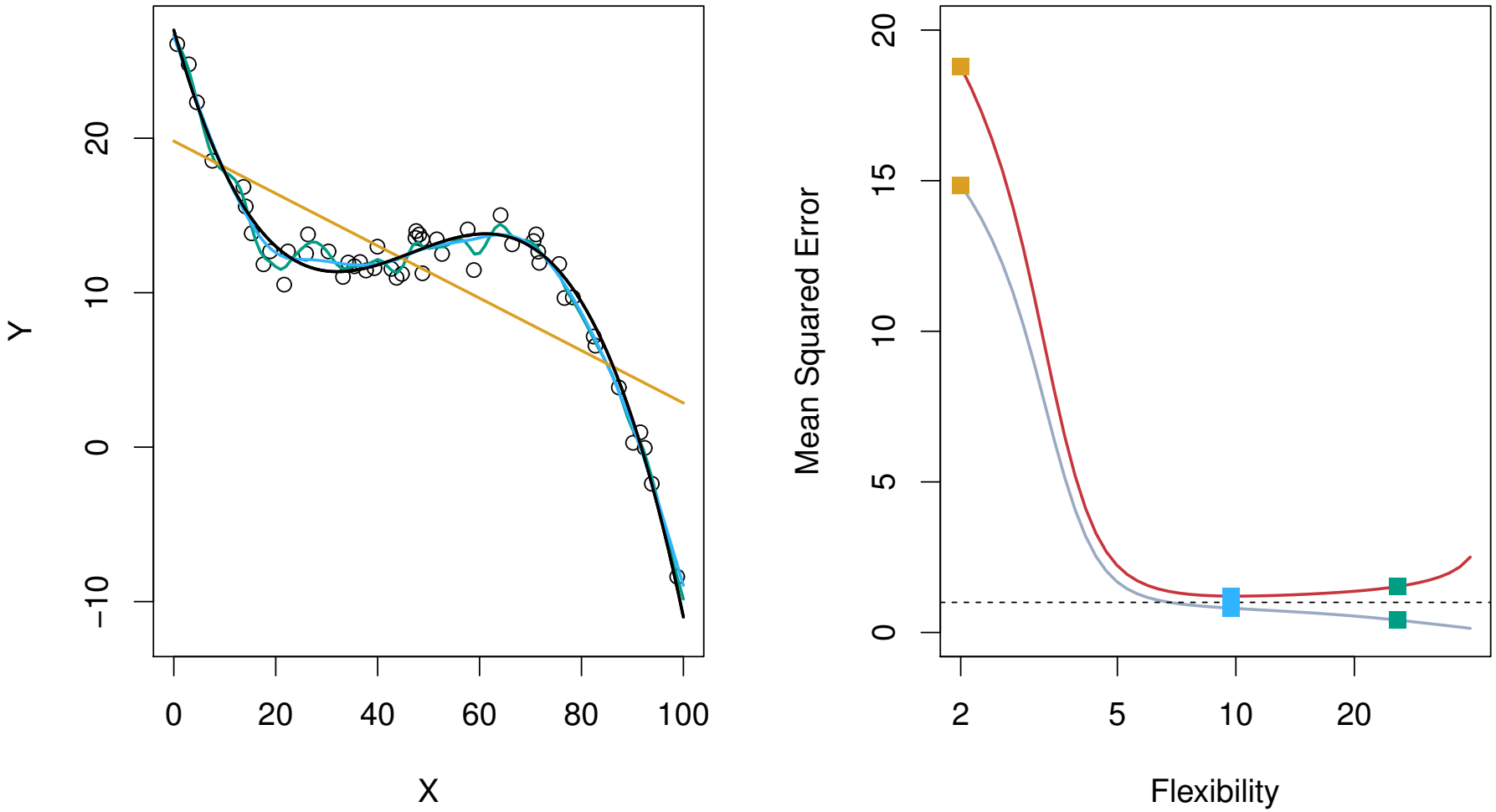
Left: true f (black), linear fit (orange), smoothing spline fits (blue & green).
Right: training error (grey), prediction error (red), Var(e) (dashed).

Assessing the Goodness of Fit



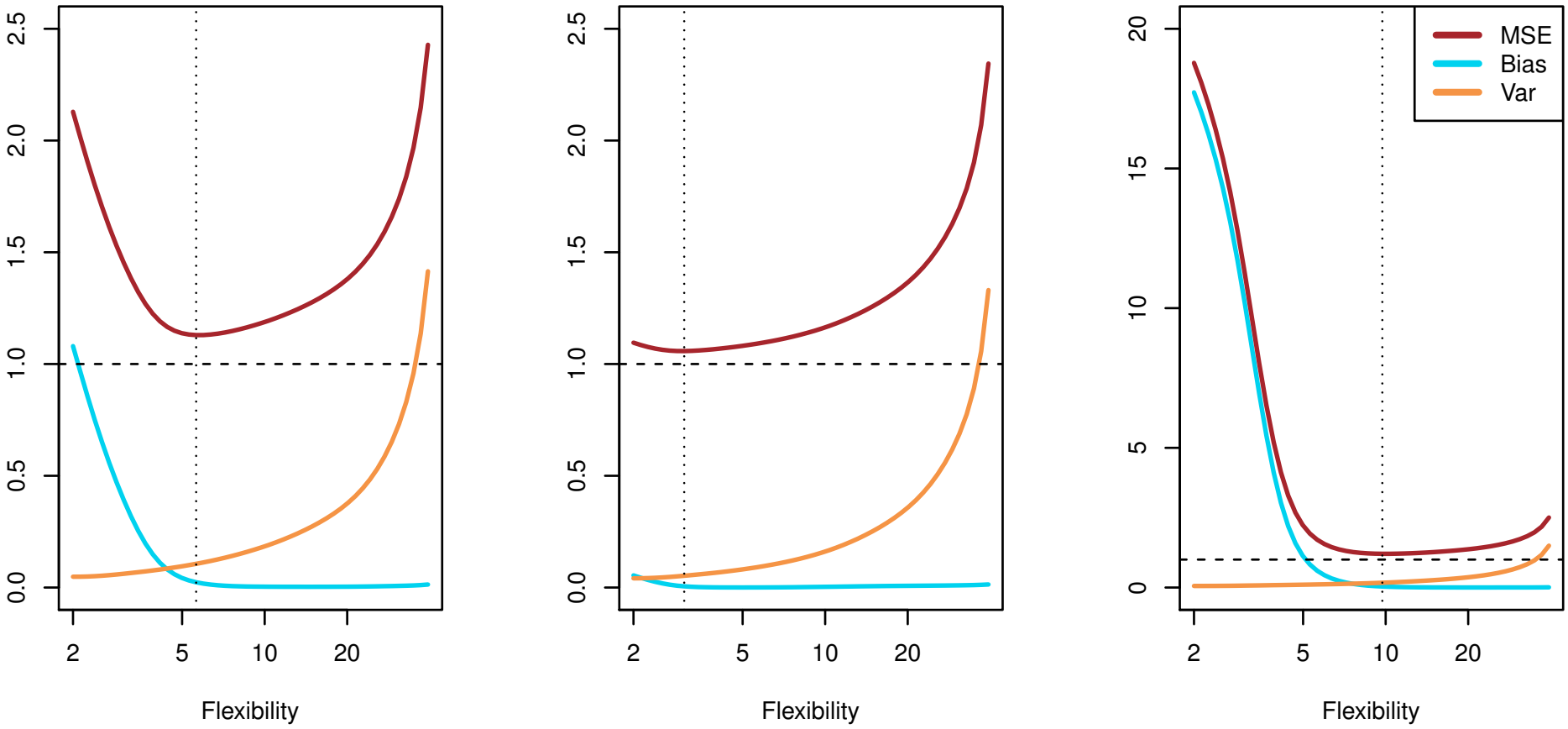
Left: true f (black), linear fit (orange), smoothing spline fits (blue & green).
Right: training error (grey), prediction error (red), Var(e) (dashed).

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

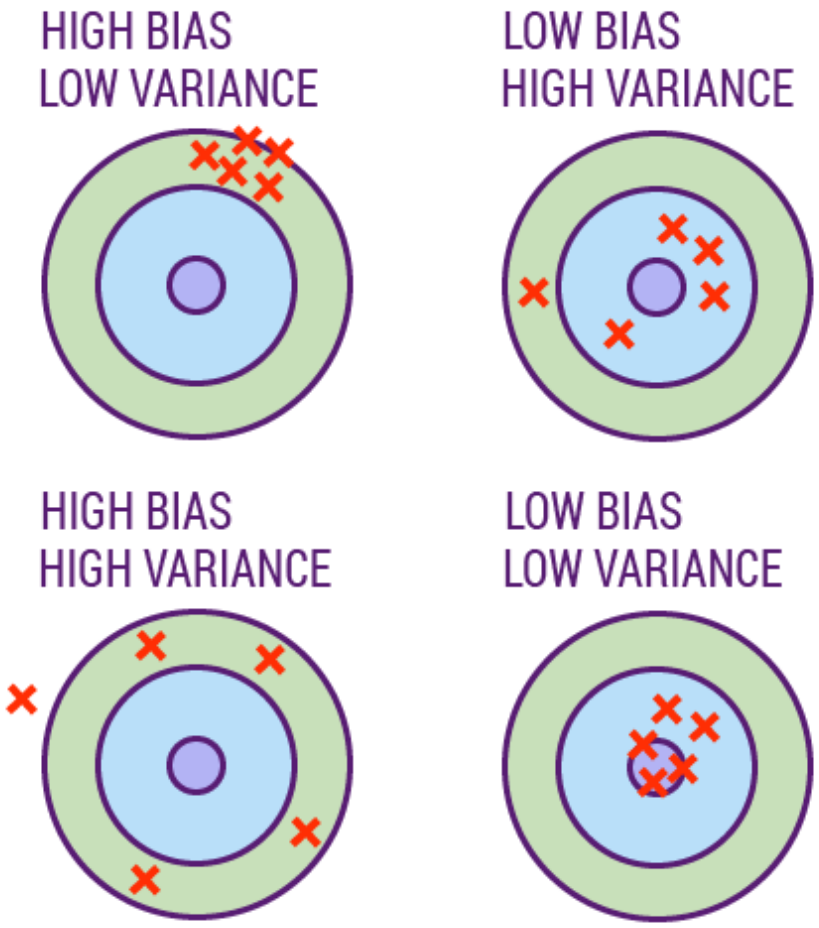
The Bias-Variance Trade-off

At a given x,

$$\mathbb{E}_{\mathcal{D}_{TR}}\left[\left(f\left(x\right)-\widehat{f}\left(x\right)\right)^2\right]=Var\left(\widehat{f}\left(x\right)\right)+\left(\text{bias}\left(\widehat{f}\left(x\right)\right)\right)^2$$

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

The Bias-Variance Trade-off



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.

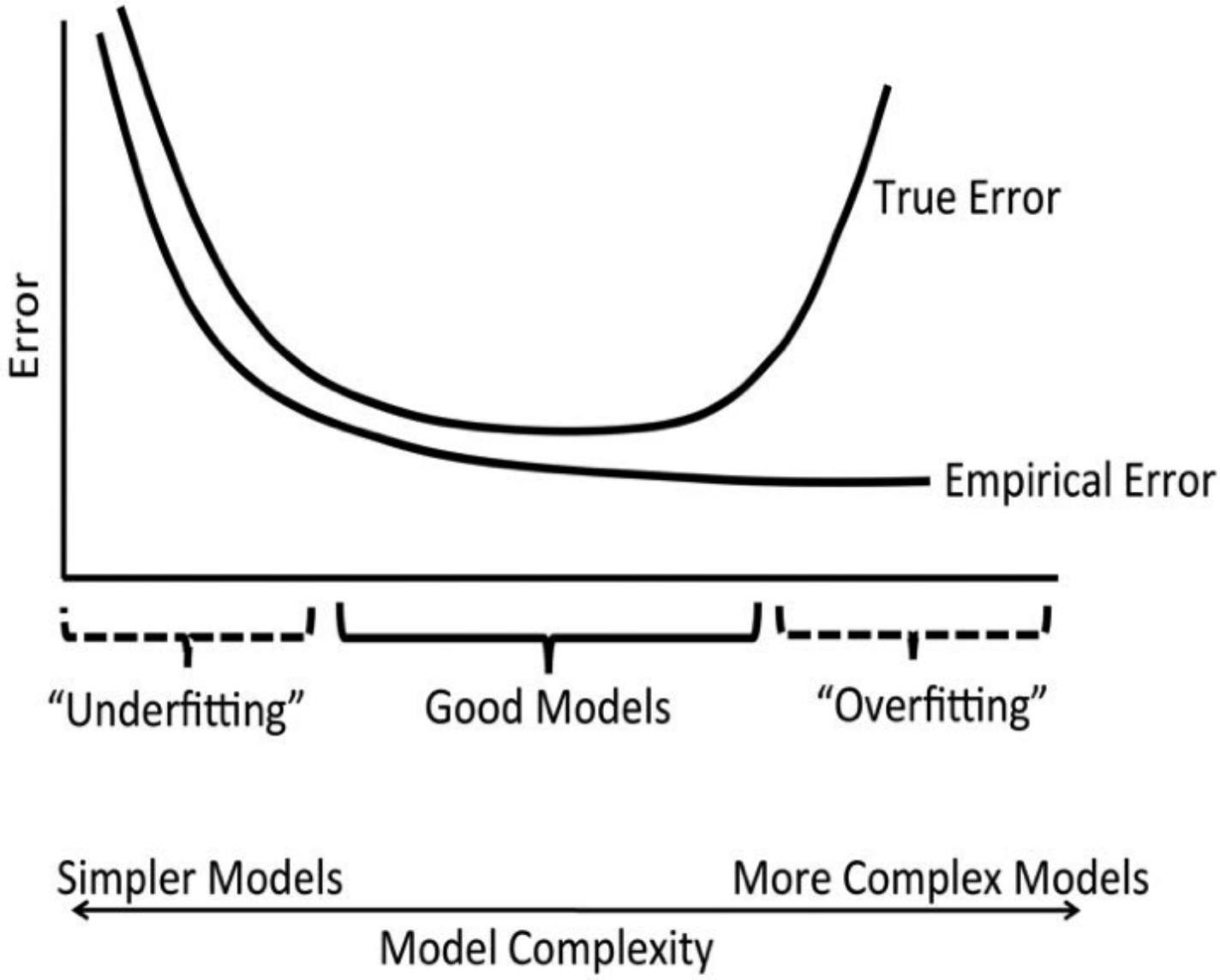
- This is due to the **bias-variance trade-off**: as model flexibility increases, the bias tends to initially decrease faster than the variance increases. Then at some point increasing flexibility has little impact on the bias but starts to significantly increase the variance.

The Bias-Variance Trade-off

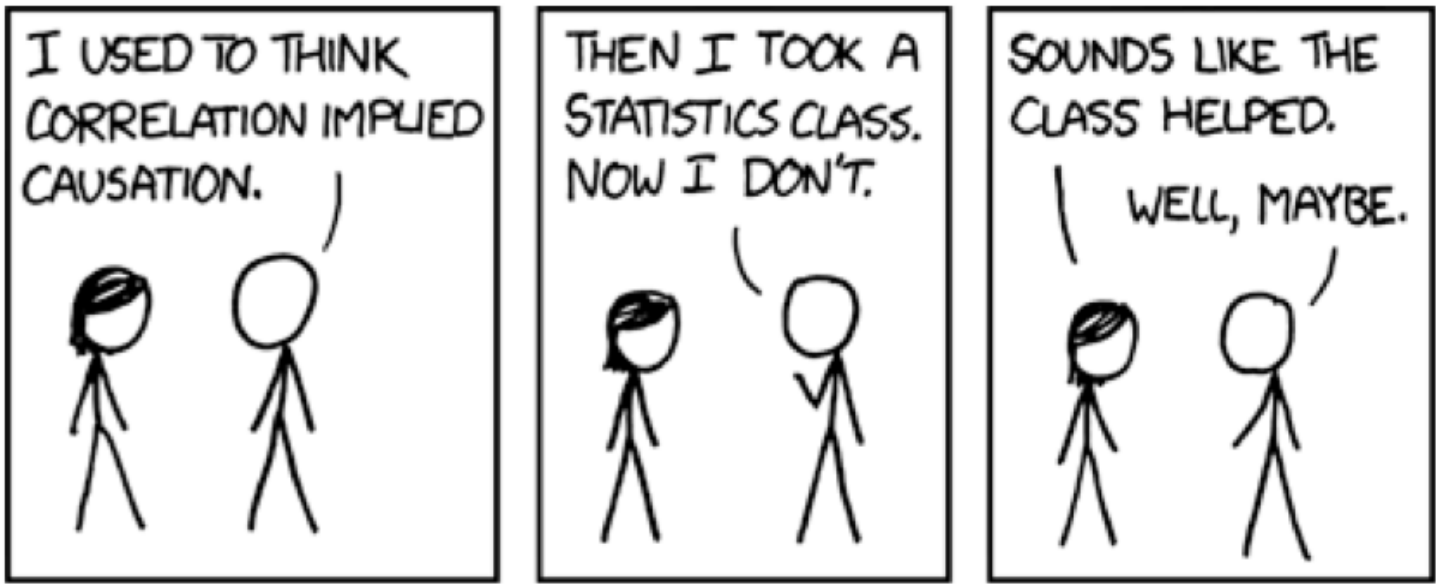
- The bias-variance trade-off is a trade-off because it is easy to have a model with extremely low bias but high variance (e.g., by drawing a curve that passes through every single training observation) or one with very low variance but high bias (e.g., by fitting a horizontal line to the data). The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



Causal Inference



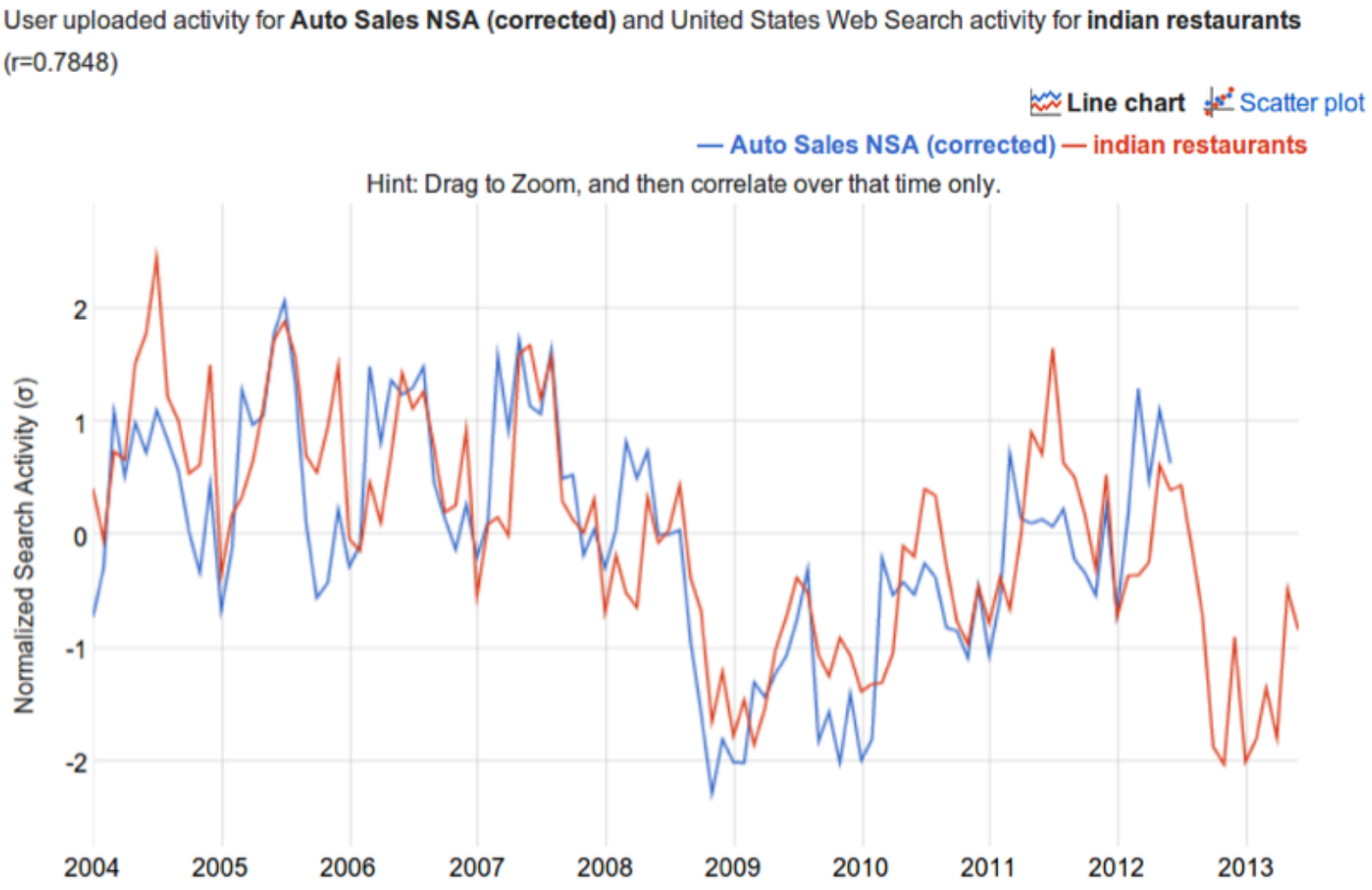
Causal Inference

- Learning the statistical relationship between x and y tells us nothing about whether there exists a causal relationship between them.

- **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



Automobile Sales and Search for Indian Restaurants

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[8] 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.

[8] 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* $x = a$?
  - $\mathbb{E}[y|\text{do}(x = a)]$[9]
- **Statistical prediction**: What will y be if I *observe* $x = a$?
  - $\mathbb{E}[y|x = a]$

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

Causal Effect Learning

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

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

Causal Effect Learning

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**)[10] is defined as

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

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

For data generated by randomized experiments, correlation implies causation.

[10] The terms "treatment effect" and "causal effect" are used interchangeably.

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

The Experimental Ideal and Its Limitations

- For many causal inference problems, RCTs are impossible or impractical to run.
  - infeasibility (e.g., monetary policy)
  - ethical reasons (e.g., smoking and lung cancer)
  - cost and duration (e.g., childhood intervention and adult outcomes)

The Experimental Ideal and Its Limitations

- Results from many RCT studies suffer from a lack of **external validity** or **inability to scale**.
  - The ATE computed from an RCT study represents the average treatment effect in the *experiment population*, which is often different from – and significantly smaller than – the *target population*[11].
  - A treatment may have very different effects when it is applied to a small RCT sample and when it is applied to a significant proportion of a large population due to equilibrium effects.

[11] Fundamentally, this problem is due to the highly heterogeneous nature of many treatment effects.

External Validity

"Psychology is the study of psychology students." – Anonymous

Vol 466|1 July 2010 nature

OPINION

Most people are not WEIRD

To understand human psychology, behavioural scientists must stop doing most of their experiments on Westerners, argue **Joseph Henrich**, **Steven J. Heine** and **Ara Norenzayan**.

A 2008 survey of the top psychology journals found that 96% of subjects were from Western, educated, industrialized, rich and democratic (WEIRD) societies – particularly American undergraduates.

Observational Studies

For observational data, correlation no longer implies causation.

Consider the following example: suppose we observe patients at two hospitals:

Hospital	Sample Size	Recovery Rate
A	1274	97%
B	569	72%

Observational Studies

Based on this observation, can we conclude that hospital A is better?

- If patients are randomly administered to hospitals – in other words, if the data come from a randomized experiment, then yes.
- In observational studies, however, it could well be that hospital B is associated with worse outcomes because it is actually better, so that people with worse health problems *choose* to visit B.

Observational Studies

- In this case, let x denote hospital choice and y denote recovery rate. Then

$$\mathbb{E}[y|x] \neq \mathbb{E}[y|\text{do}(x)]$$

: when we observe a person visiting hospital B, we would expect a lower recovery rate ($\mathbb{E}[y|x = B]$) – because she is likely sicker – than the recovery rate we would expect if we randomly assign a person to hospital B ($\mathbb{E}[y|\text{do}(x = B)]$).

- This is called **self-selection effect** or **self-selection bias**.

Observational Studies

- Self-selection is of central concern to causal inference based on observed socio-economic data generated by individual choices.
- When individuals choose their own treatments, those who choose to receive a treatment can be *systematically* different from those who choose not to. If we compare their outcomes directly, then we are comparing apples with oranges[12].

[12] Note that such self-selection effect does not exist under random assignment of hospitals because the patients administered to each hospital would be similar. Comparing their outcomes would be comparing apples with apples.

Observational Studies

- To conduct valid causal inference on the effectiveness of hospital treatment, we need to compare recovery rates of patients with the same degree of illness who visit each hospital, i.e., we need compare apples with apples.

- Let z denote patient health prior to hospital visit, then

$$\mathbb{E}[y|\text{do}(x), z] = \mathbb{E}[y|x, z]$$

Conditional on patient illness, correlation between hospital choice and recovery rate implies causation[13]!

[13] To get the overall causal effect,

$$\mathbb{E}[y|\text{do}(x)] = \mathbb{E}_z[\mathbb{E}[y|\text{do}(x), z]] = \mathbb{E}_z[\mathbb{E}[y|x, z]]$$

Observational Studies

- As the example shows, to conduct causal inference on observational studies, we need to know how the data are generated (patients choose to visit different hospitals) and why outcomes may differ among treatment groups (patients administered to different hospitals are different in degree of illness and hospitals vary in their effectiveness – the latter is the treatment effect we are interested in)[14].
- Causal inference[15] requires an understanding of the causal mechanism that generates the data[16].

[14] But wait! What if hospitals accept different health insurance plans? Suppose hospital B accepts Medicare but hospital A does not, so that hospital B has many more older patients. How does this information change our causal inference?

[15] More precisely, causal effect learning. We will later discuss causal mechanism learning – how to discover the data-generating causal mechanism in the first place.

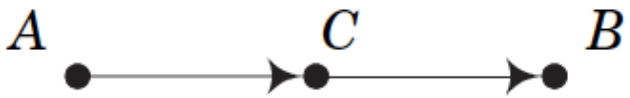
[16] As we will see, such understanding is not only necessary for observational studies but also necessary for interpreting and using experimental results. Without an understanding of – or making assumptions on – the underlying mechanism, any causal effect estimate is meaningless.

Causal Diagrams

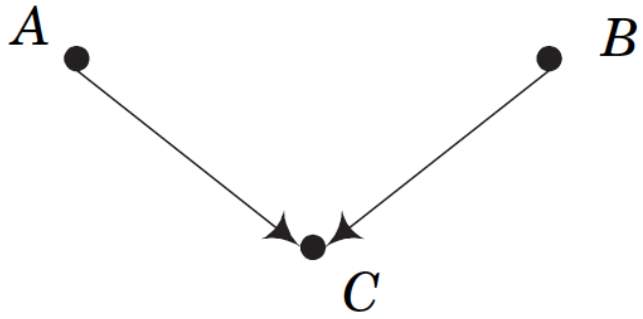
- **Causal diagrams** are graphs that represent causal relationships and therefore describe our qualitative knowledge about the causal mechanisms generating our observed data.

- In a causal diagram, the **nodes** (**vertices**) represent variables, with **directed edges** (**arrows**) representing direct causation. A sequence of connected nodes is called a **path**. The path is **causal** if all its arrows point in the same direction. Otherwise it is **noncausal**.

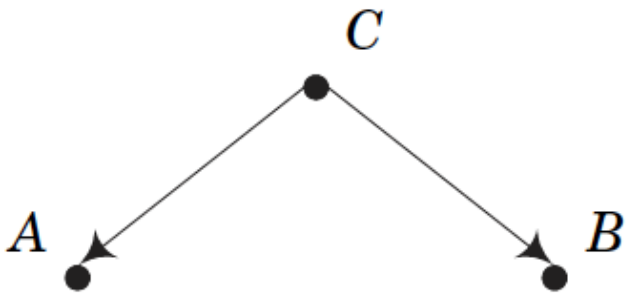
Causal Diagrams



(a) Mediation



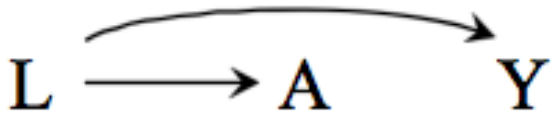
(c) Mutual causation



(b) Mutual dependence

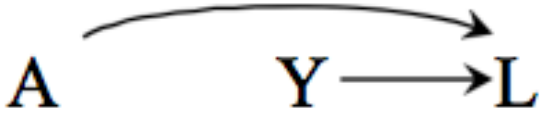
Basic patterns of causal relationships among three variables

Correlation 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.
- There exists an open path connecting A and Y: $A \leftarrow L \rightarrow Y$.
- A and Y are correlated: having information about A improves our ability to predict Y, even though A does not have a causal effect on Y.
- **Example:** A : carrying a lighter; Y : lung cancer; L : smoking

Correlation and Causation



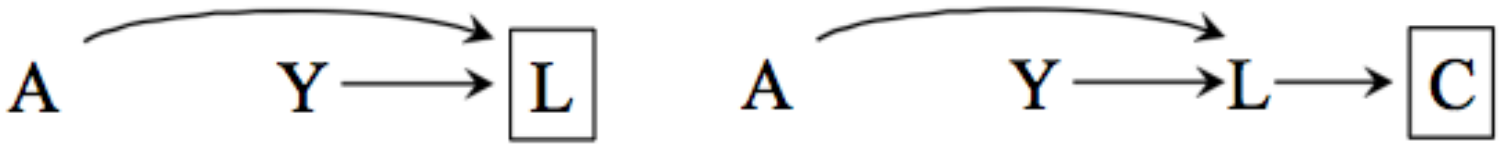
- Both A and Y have a causal effect on L. A does not have a causal effect on Y.
- L is called a **common effect** of A and Y.
- On the path $A \rightarrow L \leftarrow Y$, L is called a **collider**. The path is said to be blocked by the collider.
- A and Y are statistically independent.
- **Example:** A : family heart disease history; Y : smoking; L : heart disease

Correlation and Causation

Box indicates conditioning

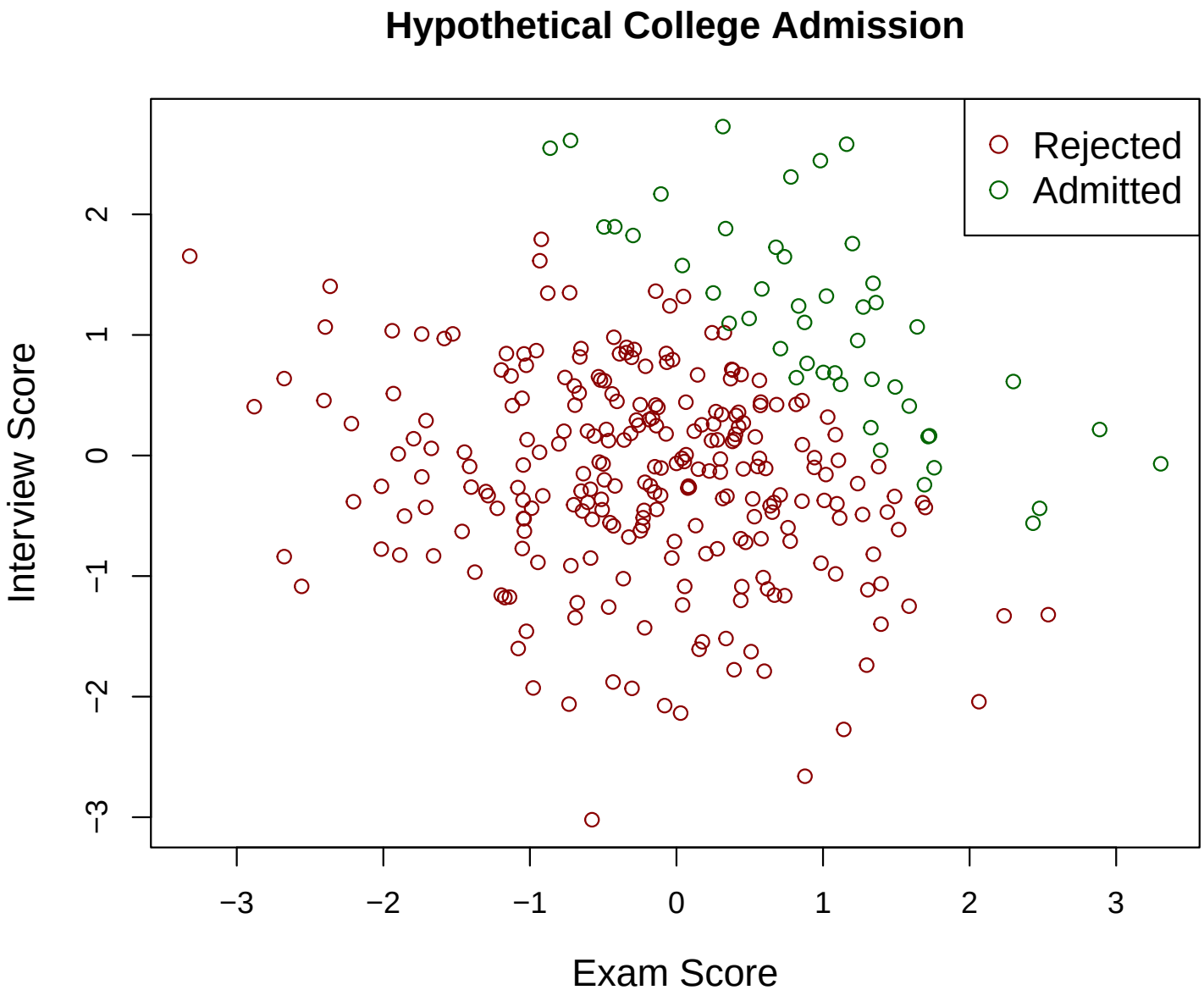
- A and Y are conditionally independent after conditioning on B and L, even though they are marginally correlated in both graphs.
- Conditioning on B and L block the paths $A \to B \to Y$ and $A \leftarrow L \to Y$.
- **Example:** (left) A : smoking; B : tar deposits in lung; Y : lung cancer

Correlation and Causation



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

Correlation and Causation



Correlation and Causation

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

1. One causes the other[17]
2. They share common causes
3. The analysis is conditioned on their common effects[18]

[17] either directly or through mediating variables.
[18] 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: when patients choose hospitals based on their illness, illness is a common cause of both their treatment (hospital) and their outcome (recovery rate), and is therefore a confounder in the analysis of the causal effect of hospital treatment.

Confounding

- A basic strategy to deal with confounding is to condition on the common causes of treatment and outcome[19] (while avoiding controlling for 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[20].

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

[20] Another way to understand this strategy: after common causes are conditioned on, to the extent that individuals who receive different treatments are still different, the differences do not affect their outcomes. Hence, when we compare different treatment groups, we would be effectively comparing apples with apples.

The Back-Door Criterion

- More generally, if we can condition on a set of variables z that block all open noncausal paths[21] between treatment x and outcome y, then the causal effect of x on y is identified[22].
  - In this case, z is said to satisfy the **back-door criterion**[23].
  - Conditioning on z makes x **exogenous** to y[24].

[21] Noncausal paths between x and y are called **back-door paths**. These are the paths that, if left open, induce correlation between x and y that is not a result of x causing y.

[22] A causal effect is **identified** if it is possible to be estimated from observed data.

[23] We also need to make sure z does not contain variables that are the common effects of x and y.

[24] x is said to be **exogenous** to y if there is no open noncausal path between the two variables (and y does not cause x). Otherwise, x is **endogenous**.

The Back-Door Criterion

Given z that satisfies the back-door criterion, we have:

$$\mathbb{E}\left[y|\text{do}(x), z\right] = \mathbb{E}\left[y|x, z\right]$$

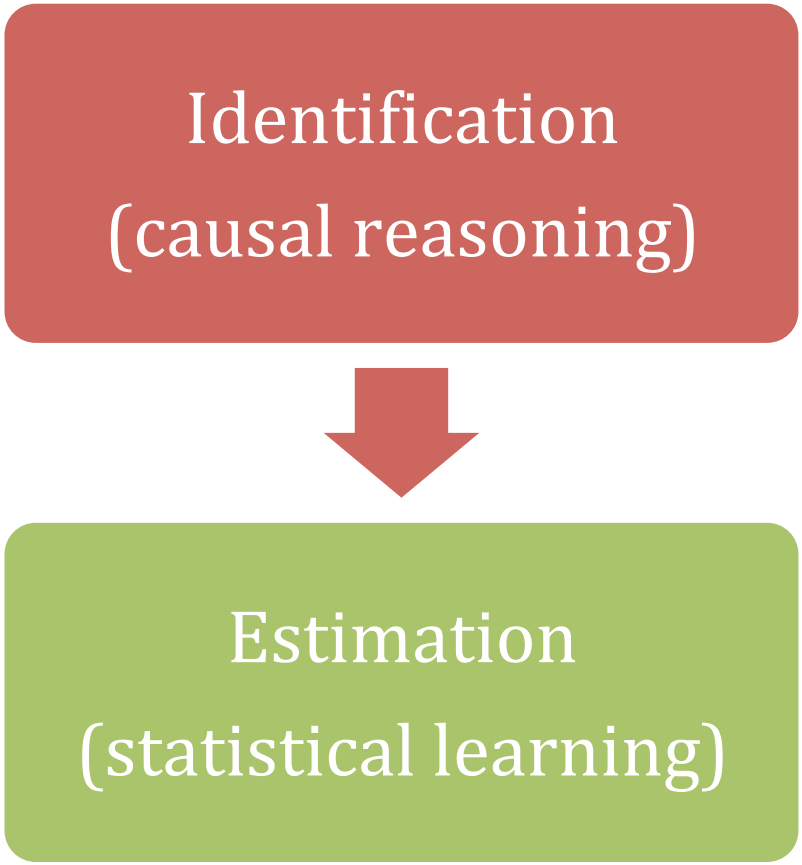
The average treatment effect[25]

$$\begin{aligned} \text{ATE} &\doteq \frac{d}{dx}\mathbb{E}\left[y|\text{do}(x)\right] \\ &= \frac{d}{dx}\mathbb{E}_z\left[\mathbb{E}\left[y|\text{do}(x), z\right]\right] \\ &= \frac{d}{dx}\mathbb{E}_z\left[\mathbb{E}\left[y|x, z\right]\right] = \mathbb{E}_z\left[\frac{\partial}{\partial x}\mathbb{E}\left[y|x, z\right]\right] \end{aligned} \tag{1}$$

[25] When $x \in \{0, 1\}$ is binary,

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

Causal Effect Learning: Two Stages



Causal Effect Learning: Two Stages

- Once we have established identification based on causal reasoning, we can estimate the causal effect of interest using statistical models.

- Causal effect learning is therefore a two-stage process: in the first stage we determine what correlations in the data can tell us about causation (causal reasoning). In the second stage, we estimate the correlations from data (statistical learning).

Causal Effect Learning: Two Stages

- For example, suppose we have established that a set of observed variables z satisfies the back-door criterion, then according to (1), estimation of the ATE requires estimation of $\mathbb{E}[y|x,z]$.

- To estimate $\mathbb{E}[y|x,z]$ from data, we can use a variety of statistical models:
  - parametric or nonparametric
  - linear or non-linear

Causal Effect Learning: Two Stages

For simplicity, let

$$\mathbb{E}[y|x,z] = \beta_0 + \beta_1 x + \beta_2 z$$

Then

$$\widehat{\mathsf{ATE}} = \mathbb{E}_z\left[\frac{\partial}{\partial x}\widehat{\mathbb{E}}[y|x,z]\right] = \widehat{\beta}_1$$