ML@LSE Bootcamp 1 - Introduction

Springer Texts in Statistics

Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani

An Introduction to Statistical Learning

with Applications in R

Springer

What is Data?

Plan:

A- Data Structure

---

1. **What is Data?**

a. Intuitive examples
b. Datapoint as a vector

2. **Datasets**

a. Dataset, number of observations vs dimensionality
b. Large n, large p and associated techniques

B- Machine Learning is all about finding patterns

---

1. **Finding patterns in Data: Supervised vs. Unsupervised**

a. What are patterns: use your brain's intuition!
b. Two kinds of patterns, two Machine Learning fields: Supervised vs. Unsupervised.

2. More on Supervised Learning

a. The learning function and training data
b. Regression vs. Classification

C- Assessing the model accuracy

---

1. **Measuring the quality of fit to data**

a. Risk of a learning function
b. Empirical Risk: MSE and MER

2. **The Bias variance trade-off**

a. Overfitting vs. Generalization
b. Cross-Validation

Data = information?

Examples of various Datasets:

- Quarterly GDP of the USA from 1947 to 2018 (Time Series)
- Boston Housing Data (cross sectional)

FRED Graph Observations	
Federal Reserve Economic Data	
Link: https://fred.stlouisfed.org	
Help: https://fred.stlouisfed.org/help-faq	
Economic Research Division	
Federal Reserve Bank of St. Louis	
GDP	Gross Domestic Product, Billions of Dollars, Quarterly, Seasonally Adjusted Annual Rate
Frequency: Quarterly	
observation_date	GDP
1947-01-01	243.164
1947-04-01	245.968
1947-07-01	249.585
1947-10-01	259.745
1948-01-01	265.742
1948-04-01	272.567
1948-07-01	279.196
1948-10-01	280.366
1949-01-01	275.034
1949-04-01	271.351
1949-07-01	272.889
1949-10-01	270.627
1950-01-01	280.828
1950-04-01	290.383
1950-07-01	308.153
1950-10-01	319.945
1951-01-01	336.000
1951-04-01	344.090
1951-07-01	351.385
1951-10-01	356.178
1952-01-01	359.820
1952-04-01	361.030
1952-07-01	367.701
1952-10-01	380.812
1953-01-01	387.980
1953-04-01	391.749
1953-07-01	391.171
1953-10-01	385.970

What is Data?

Plan:

A- Data Structure

1. **What is Data?**

a. Intuitive examples
b. Datapoint as a vector

2. **Datasets**

a. Dataset, number of observations vs dimensionality
b. Large n, large p and associated techniques

B- Machine Learning is all about finding patterns

1. **Finding patterns in Data: Supervised vs. Unsupervised**

a. What are patterns: use your brain's intuition!
b. Two kinds of patterns, two Machine Learning fields: Supervised vs. Unsupervised.

2. More on Supervised Learning

a. The learning function and training data
b. Regression vs. Classification

C- Assessing the model accuracy

1. **Measuring the quality of fit to data**

a. Risk of a learning function
b. Empirical Risk: MSE and MER

2. **The Bias variance trade-off**

a. Overfitting vs. Generalization
b. Cross-Validation

Data = information?

Examples of various Datasets:

- Quarterly GDP of the USA from 1947 to 2018 (Time Series)
- Boston Housing Data (cross sectional)

	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
1	0.00632	18.0	2.31	0	0.5380	6.575	65.2	4.0900	1	296	15.3	396.90	4.98	24.0
2	0.02731	0.0	7.07	0	0.4690	6.421	78.9	4.9671	2	242	17.8	396.90	9.14	21.6
3	0.02729	0.0	7.07	0	0.4690	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
4	0.03237	0.0	2.18	0	0.4580	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
5	0.06905	0.0	2.18	0	0.4580	7.147	54.2	6.0622	3	222	18.7	396.90	5.33	36.2
6	0.02985	0.0	2.18	0	0.4580	6.430	58.7	6.0622	3	222	18.7	394.12	5.21	28.7
7	0.08829	12.5	7.87	0	0.5240	6.012	66.6	5.5605	5	311	15.2	395.60	12.43	22.9
8	0.14455	12.5	7.87	0	0.5240	6.172	96.1	5.9505	5	311	15.2	396.90	19.15	27.1
9	0.21124	12.5	7.87	0	0.5240	5.631	100.0	6.0821	5	311	15.2	386.63	29.93	16.5
10	0.17004	12.5	7.87	0	0.5240	6.004	85.9	6.5921	5	311	15.2	386.71	17.10	18.9
11	0.22489	12.5	7.87	0	0.5240	6.377	94.3	6.3467	5	311	15.2	392.52	20.45	15.0
12	0.11747	12.5	7.87	0	0.5240	6.009	82.9	6.2267	5	311	15.2	396.90	13.27	18.9
13	0.09378	12.5	7.87	0	0.5240	5.889	39.0	5.4509	5	311	15.2	390.50	15.71	21.7
14	0.62976	0.0	8.14	0	0.5380	5.949	61.8	4.7075	4	307	21.0	396.90	8.26	20.4
15	0.63796	0.0	8.14	0	0.5380	6.096	84.5	4.4619	4	307	21.0	380.02	10.26	18.2
16	0.62739	0.0	8.14	0	0.5380	5.834	56.5	4.4986	4	307	21.0	395.62	8.47	19.9
17	1.05393	0.0	8.14	0	0.5380	5.935	29.3	4.4986	4	307	21.0	386.85	6.58	23.1
18	0.78420	0.0	8.14	0	0.5380	5.990	81.7	4.2579	4	307	21.0	386.75	14.67	17.5
19	0.80271	0.0	8.14	0	0.5380	5.456	36.6	3.7965	4	307	21.0	288.99	11.69	20.2
20	0.72580	0.0	8.14	0	0.5380	5.727	69.5	3.7965	4	307	21.0	390.95	11.28	18.2
21	1.25179	0.0	8.14	0	0.5380	5.570	98.1	3.7979	4	307	21.0	376.57	21.02	13.6
22	0.85204	0.0	8.14	0	0.5380	5.965	89.2	4.0123	4	307	21.0	392.53	13.83	19.6
23	1.23247	0.0	8.14	0	0.5380	6.142	91.7	3.9769	4	307	21.0	396.90	18.72	15.2
24	0.98843	0.0	8.14	0	0.5380	5.813	100.0	4.0952	4	307	21.0	394.54	19.88	14.5

Datapoint as a vector

A *datapoint* is a vector:

- Individual X → (Number of years at school of individual X, Income of individual X)
- District X → (Median housing price in X, Crime rate in X, Average number of rooms per dwelling in X, Proportion of residential land, etc…)

More Generally:

$\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})$

Big Data

Plan:

A- Data Structure

1. What is Data?

a. Intuitive examples
b. Datapoint as a vector

2. Datasets

a. Dataset, number of observations vs dimensionality
b. Large n, large p and associated techniques

B- Machine Learning is all about finding patterns

1. Finding patterns in Data: Supervised vs. Unsupervised

a. What are patterns: use your brain's intuition!
b. Two kinds of patterns, two Machine Learning fields: Supervised vs. Unsupervised.

2. More on Supervised Learning

a. The learning function and training data
b. Regression vs. Classification

C- Assessing the model accuracy

1. Measuring the quality of fit to data

a. Risk of a learning function
b. Empirical Risk: MSE and MER

2. The Bias variance trade-off

a. Overfitting vs. Generalization
b. Cross-Validation

Dataset: A set of points (vectors in a vector space) whose coordinates are associated with a specific feature (age, height, weight etc…).

Two important characteristics of a dataset:
- n: the *number of observations* (the number of data points in our dataset)
- p: the number of features, i.e. the *dimensionality.*

- Big n: Machine Learning techniques (cloud computing)
- Big p: High dimensional statistics (when p>n)

Data can also be *unstructured*: image recognition, sentiment analysis, natural language, processing etc...

Machine Learning is all about finding patterns

Plan:

A- Data Structure

---

1. What is Data?

a. Intuitive examples
b. Datapoint as a vector

2. Datasets

a. Dataset, number of observations vs dimensionality
b. Large n, large p and associated techniques

B- Machine Learning is all about finding patterns

---

1. Finding patterns in Data: Supervised vs. Unsupervised

a. What are patterns: use your brain's intuition!
b. Two kinds of patterns, two Machine Learning fields: Supervised vs. Unsupervised.

2. More on Supervised Learning

a. The learning function and training data
b. Regression vs. Classification

C- Assessing the model accuracy

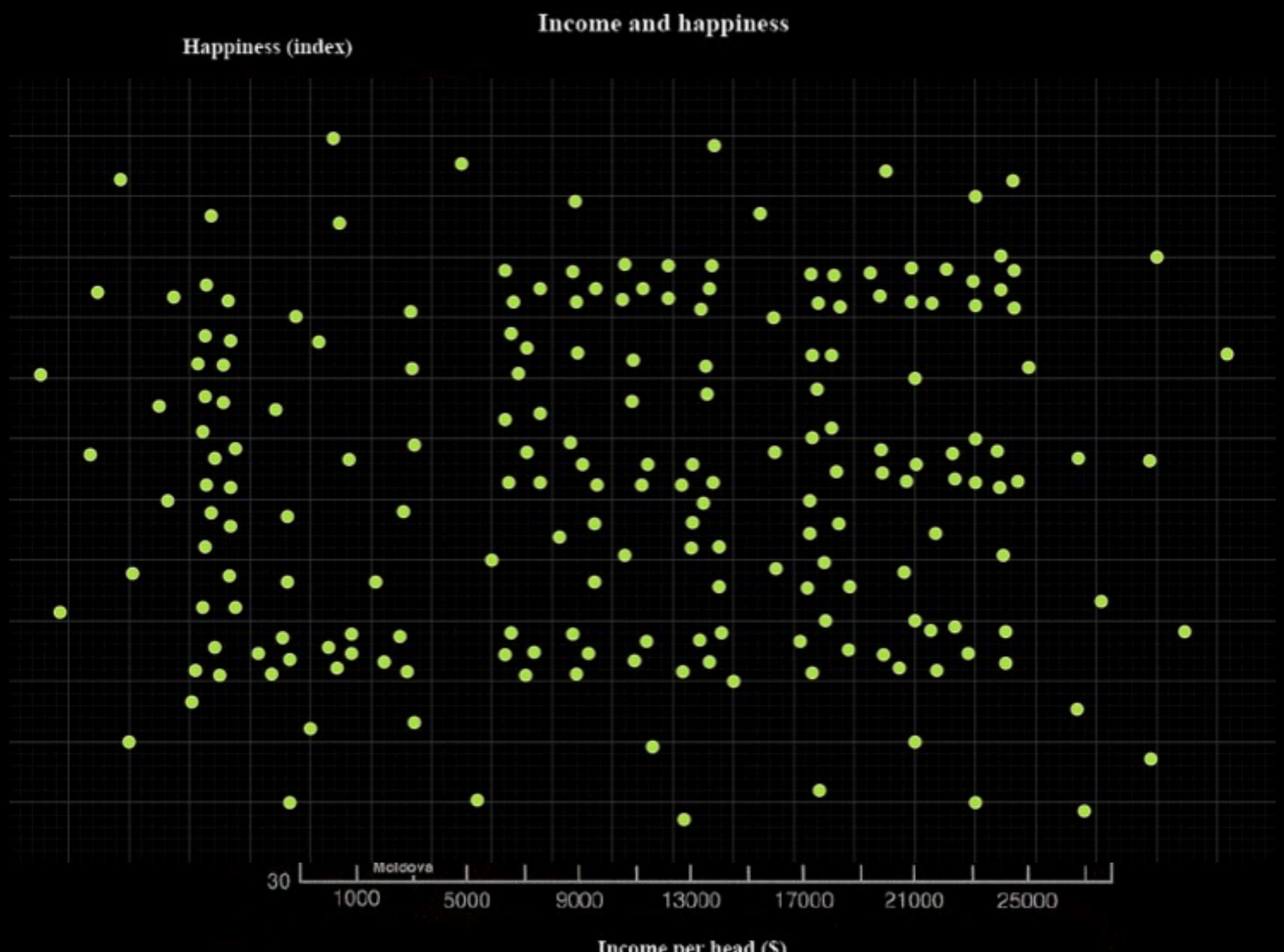
---

1. Measuring the quality of fit to data

a. Risk of a learning function
b. Empirical Risk: MSE and MER

2. The Bias variance trade-off

a. Overfitting vs. Generalization
b. Cross-Validation



Source: Inglehart and Klingemann (2000), Figure 7.2 and Table 7.1. Latest year (all in 1990s).

Supervised vs. Unsupervised learning

Plan:

A- Data Structure

---

1. **What is Data?**

a. Intuitive examples
b. Datapoint as a vector

2. **Datasets**

a. Dataset, number of observations vs dimensionality
b. Large n, large p and associated techniques

B- Machine Learning is all about finding patterns

---

1. **Finding patterns in Data: Supervised vs. Unsupervised**

a. What are patterns: use your brain's intuition!
b. Two kinds of patterns, two Machine Learning fields: Supervised vs. Unsupervised.

2. More on Supervised Learning

a. The learning function and training data
b. Regression vs. Classification

C- Assessing the model accuracy

---

1. **Measuring the quality of fit to data**

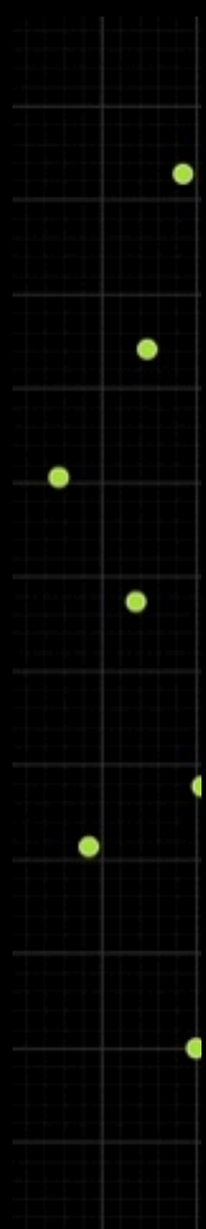
a. Risk of a learning function
b. Empirical Risk: MSE and MER

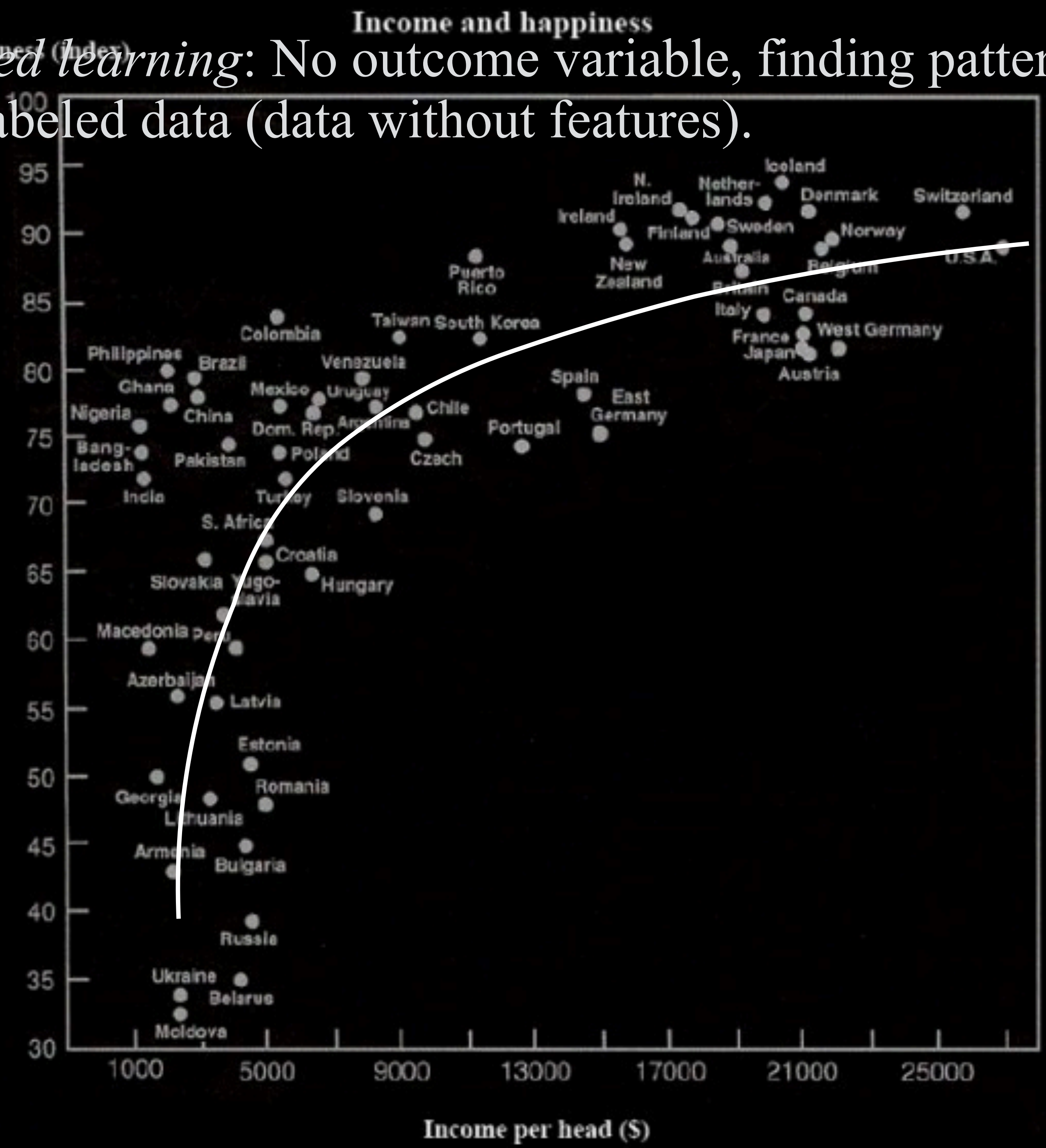
2. **The Bias variance trade-off**

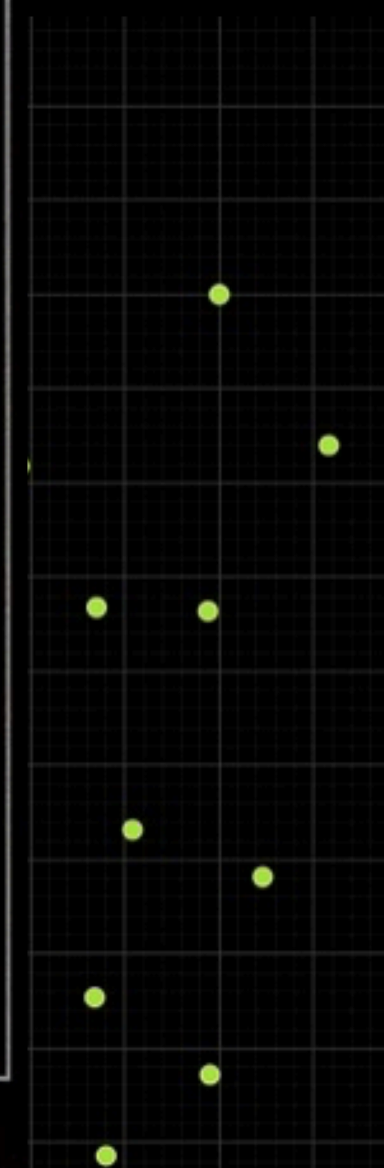
a. Overfitting vs. Generalization
b. Cross-Validation

Supervised learning: finding the relationship between some input variables and one outcome variable.

Unsupervised learning: No outcome variable, finding patterns to describe usually unlabeled data (data without features).







Source: Inglehart and Klingemann (2000), Figure 7.2 and Table 7.1. Latest year (all in 1990s).

Supervised learning: the learning function

Y_i: *Dependent variable*, outcome variable, response variable (Happiness; House price; Income), Y_i a random variable.

$\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})$, the set of *input variables*, features, regressors (GDP per capita of the five previous years; Crime Rate, Number of trees; Education), $\mathbf{X}_i$ is a RV.

We model the relationship between the input variables and the output variables as follows:

$$Y_i = f(\mathbf{X}_i) + \varepsilon_i$$

Where f is the function to be estimated and ε represents our ignorance, (random variable with mean 0 and finite variance).

The estimated function of f is written $\hat{f}$.

We use the *training data* $(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)$ to estimate $\hat{f}$.

The input and output variables are random, the training data is observed.

Supervised learning: the learning function

Plan:

A- Data Structure

---

1. **What is Data?**

a. Intuitive examples
b. Datapoint as a vector

2. **Datasets**

a. Dataset, number of observations vs dimensionality
b. Large n, large p and associated techniques

B- Machine Learning is all about finding patterns

---

1. **Finding patterns in Data: Supervised vs. Unsupervised**

a. What are patterns: use your brain's intuition!
b. Two kinds of patterns, two Machine Learning fields: Supervised vs. Unsupervised.

2. More on Supervised Learning

a. The learning function and training data
b. Regression vs. Classification

C- Assessing the model accuracy

---

1. **Measuring the quality of fit to data**

a. Risk of a learning function
b. Empirical Risk: MSE and MER

2. **The Bias variance trade-off**

a. Overfitting vs. Generalization
b. Cross-Validation

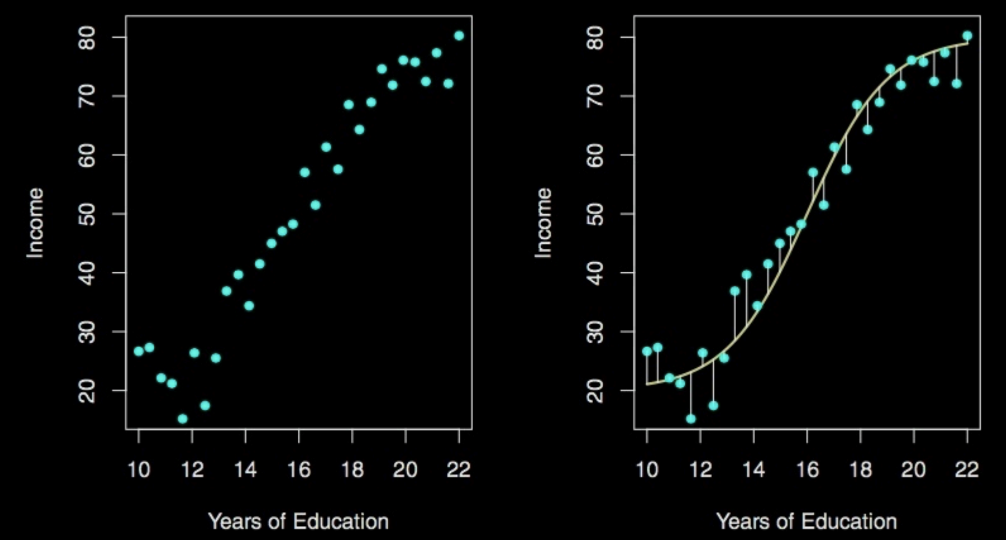
Intuitively, what desirable properties should f have?

We'd like f to have some good prediction accuracy, that is, given a set of features $\mathbf{x}_i$, we want $\hat{f}(\mathbf{x}_i)$ to be close to the truth y_i on average.

Hence we want to find $\hat{f}$ so that our model minimizes the total discrepancies between the *estimated* value of the y_is, i.e $\hat{f}(\mathbf{x}_i)$, and the *actual* y_is.

The discrepancy is usually measured as the sum of squares of the difference between each $\hat{f}(\mathbf{x}_i)$ and y_i:

$\sum (\hat{f}(\mathbf{x}_i) - y_i)^2$ *(more on this later)*



This is called the residual sum of squares (RSS)

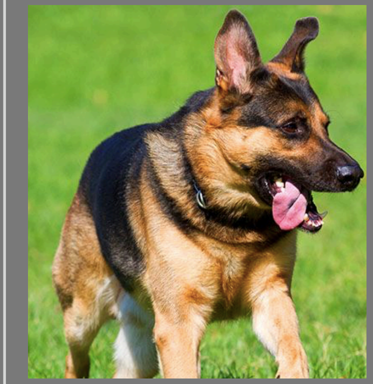
$f(\mathbf{X}_i) - Y_i = \varepsilon$ is the *irreducible error*, even if we knew the true f, we would still make some errors in prediction.

Supervised learning: regression vs. classification

Depending on the nature of the outcome variable, we're either talking about *Regression* or *Classification*:

- Regression: Y is continuous
- Classification: Y is categorical

Examples:

Input				
Output	A cat	Not a cat	A cat	Not a cat

Measuring the model accuracy: Risk of a learning function

Plan:

A- Data Structure

1. What is Data?

a. Intuitive examples
b. Datapoint as a vector

2. Datasets

a. Dataset, number of observations vs dimensionality
b. Large n, large p and associated techniques

B- Machine Learning is all about finding patterns

1. Finding patterns in Data: Supervised vs. Unsupervised

a. What are patterns: use your brain's intuition!
b. Two kinds of patterns, two Machine Learning fields: Supervised vs. Unsupervised.

2. More on Supervised Learning

a. The learning function and training data
b. Regression vs. Classification

C- Assessing the model accuracy

1. Measuring the quality of fit to data

a. Risk of a learning function
b. Empirical Risk: MSE and MER

2. The Bias variance trade-off

a. Overfitting vs. Generalization
b. Cross-Validation

How to quantify the prediction accuracy of our model?

By introducing the *loss function*:

- For regression, we use the l_2 loss function:

$$L(Y_i, f(X_i)) = (Y_i - f(X_i))^2$$

- For classification, we use the 0-1 loss function:

$$L(Y_i, f(X_i)) = I(Y_i \neq f(X_i))$$

Note that L is a function of random variables, hence it is a random variable itself! Therefore, we can define its expected value, which is called the *risk*:

$$R(f) = E[L(Y_i, f(\mathbf{X_i})]$$

Our goal will be to find f such that the risk is minimized. As we saw before, we'll use the training data to estimate such an optimal f.

Measuring the model accuracy: Empirical Risk

The empirical analogue of the risk, the *empirical risk*, given by:

$$R_n(f) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i)).$$

- For a regression, we use the *Mean Square Error* (MSE):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2.$$

- For classification, we use the *Misclassification Error Rate* (MER):

$$MER = \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{f}(x_i)).$$

Measuring the model accuracy: Risk of a learning function

It's important not to lose sight of our goal: prediction! What we really care about is how well our function predicts data out of our sample (Who cares about predicting what we already know!). This new data is called the *testing data*.

Reminder: training data = data that we have, testing data = new data.

We aim to find the function giving the lowest **test** MSE or MER, for regression and classification problems respectively, rather than the lowest training empirical risk.

We'll see next that there is a trade-off between the training MSE/MER and the test MSE/MER.

Measuring the model accuracy: Flexibility vs interpretability

Plan:

A- Data Structure

1. **What is Data?**

a. Intuitive examples
b. Datapoint as a vector

2. **Datasets**

a. Dataset, number of observations vs dimensionality
b. Large n, large p and associated techniques

B- Machine Learning is all about finding patterns

1. **Finding patterns in Data: Supervised vs. Unsupervised**

a. What are patterns: use your brain's intuition!
b. Two kinds of patterns, two Machine Learning fields: Supervised vs. Unsupervised.

2. More on Supervised Learning

a. The learning function and training data
b. Regression vs. Classification

C- Assessing the model accuracy

1. **Measuring the quality of fit to data**

a. Risk of a learning function
b. Empirical Risk: MSE and MER

2. **The Bias variance trade-off**

a. Overfitting vs. Generalization
b. Cross-Validation

What form should f take? Linear, polynomial, splines?

To understand better, let's look at an animation

Measuring the model accuracy: Bias-variance trade-off

Plan:

A- Data Structure

1. **What is Data?**
   a. Intuitive examples
   b. Datapoint as a vector
2. **Datasets**
   a. Dataset, number of observations vs dimensionality
   b. Large n, large p and associated techniques

B- Machine Learning is all about finding patterns

1. **Finding patterns in Data: Supervised vs. Unsupervised**
   a. What are patterns: use your brain's intuition!
   b. Two kinds of patterns, two Machine Learning fields: Supervised vs. Unsupervised.

2. More on Supervised Learning
   a. The learning function and training data
   b. Regression vs. Classification

C- Assessing the model accuracy

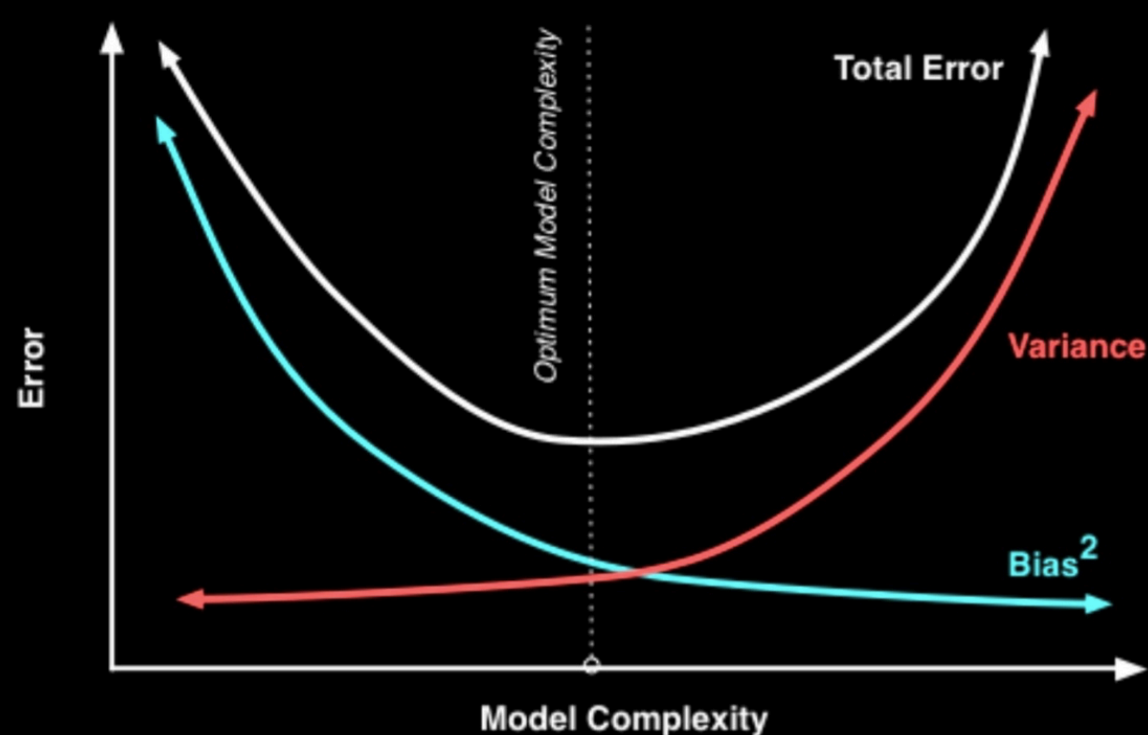
1. **Measuring the quality of fit to data**
   a. Risk of a learning function
   b. Empirical Risk: MSE and MER
2. **The Bias variance trade-off**
   a. Overfitting vs. Generalization
   b. Cross-Validation

Suppose we fit a model on our training data, then we can decompose the expected test MSE in the following way, for an arbitrary x_0:

$$E(y_0 - \hat{f}(x_0))^2 = Var(\hat{f}(x_0)) + [Bias(\hat{f}(x_0))]^2 + Var(\epsilon_0)$$

where $Bias(\hat{f}(x_0)) = E[\hat{f}(x_0)] - f(x_0)$

As the flexibility of $\hat{f}$ increases, its variance increases and its bias decreases. This is the *bias variance trade-off* that we previously explored.



Hence to minimize $E(y_0 - \hat{f}(x_0))^2$ (the expected test MSE), we must find
a method that yields both a low variance and a low bias: this is the most
important trade-off in machine learning. The minimal MSE offers the best compromise.

Cross-validation

Plan:

A- Data Structure

1. **What is Data?**
   a. Intuitive examples
   b. Datapoint as a vector
2. **Datasets**
   a. Dataset, number of observations vs dimensionality
   b. Large n, large p and associated techniques

B- Machine Learning is all about finding patterns

1. **Finding patterns in Data: Supervised vs. Unsupervised**
   a. What are patterns: use your brain's intuition!
   b. Two kinds of patterns, two Machine Learning fields: Supervised vs. Unsupervised.
2. **More on Supervised Learning**
   a. The learning function and training data
   b. Regression vs. Classification

C- Assessing the model accuracy

1. **Measuring the quality of fit to data**
   a. Risk of a learning function
   b. Empirical Risk: MSE and MER
2. **The Bias variance trade-off**
   a. Overfitting vs. Generalization
   b. Cross-Validation

If we only have training data, how can we say anything about the *test* MSE?

➢ Use only part of the data as a training data, and use the remaining of the sample as the test data. This is called Cross-validation.

K-fold cross validation:

Thanks for coming!