

An aerial photograph of a city, likely San Francisco, showing a dense urban grid with buildings and streets. A large, semi-transparent blue rectangular box is overlaid on the top half of the image, containing the title text.

# Introduction to Machine Learning

Jonathan Hersh, PhD (Chapman Argyros School of Business)

8/06/2020

# Outline

## 1. What is Machine Learning?

- Machine learning versus econometrics

## 2. Why Machine Learning at Central Banks??

- Big data requires it
- Non-linear relationships
- Better forecasts/econometrics
- Anomaly detection

## 3. Some Basic Machine Learning Concepts

- Supervised vs Unsupervised learning
- Testing/Training Sets
- Bias-Variance Tradeoff

# 1. What is Machine Learning?



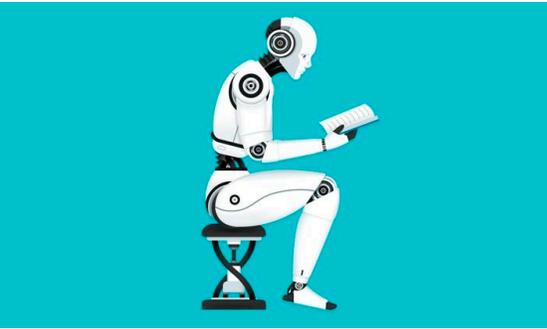
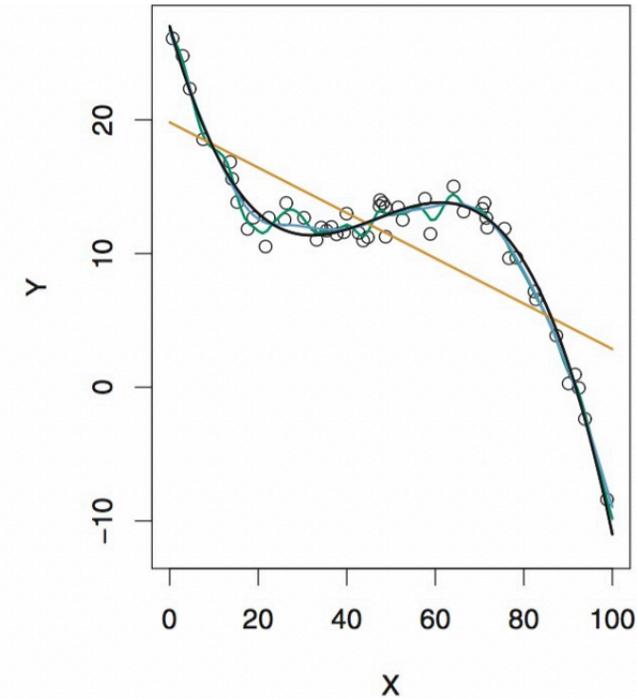
# Public Conception of Machine Learning

# Reality (90% of the time)

Target or Output

Input data

$$\hat{y} = \hat{f}(x)$$



# Machine Learning Versus Econometrics

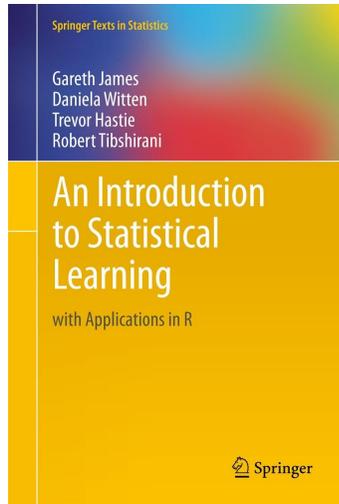
- **Machine Learning**

- Developed to solve problems in computer science
- Prediction/classification
- Desire: goodness of fit
  
- Huge Datasets! (Terabytes)  
Thousands of variables!
- Whatever works

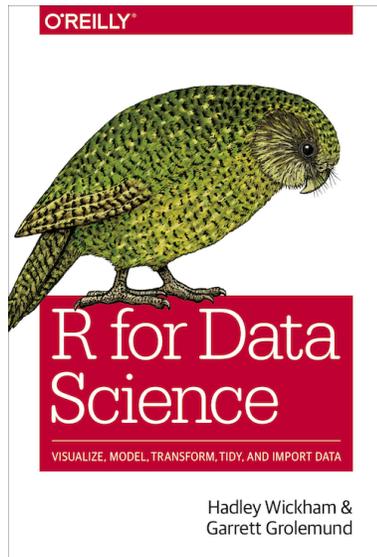
- **Econometrics**

- Developed to solve problems in economics
- Explicitly testing a theory
- “Statistical significance” more important than model fit
  
- Small datasets  
Few dozen variables
- “It works in practice, but what about theory?”

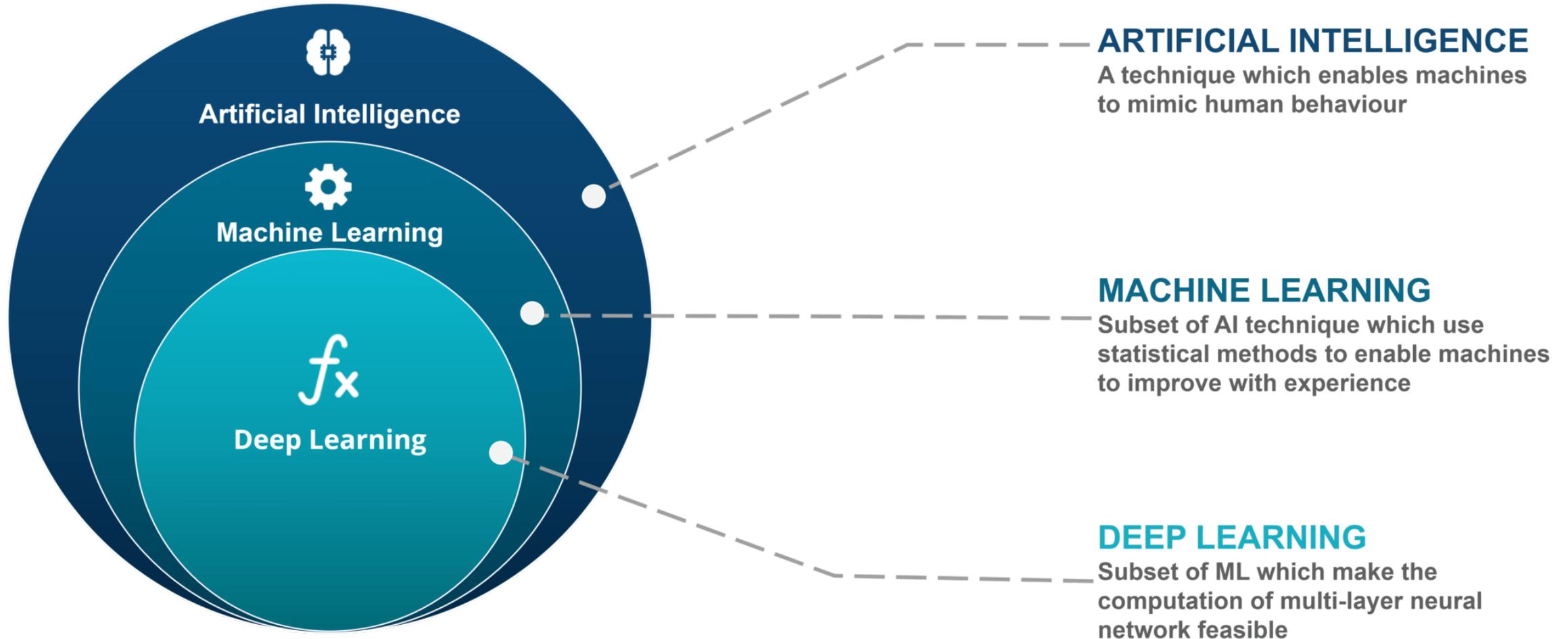
# Today – Brief Introduction to Machine Learning



- **Cross-Validation [Chapter 2 ISLR]**
- **Ridge Regression [Chapter 6 ISLR]**
- **Lasso Regression [Chapter 6 ISLR]**
- **Decision Trees [Chapter 8 ISLR]**
- **Introduction to R [R for Data Science]**



# Machine Learning Versus Artificial Intelligence



# Recent Papers on Applications of Machine Learning at Central Banks



BANK OF ENGLAND

Staff Working Paper No. 674

## Machine learning at central banks

Chiranjit Chakraborty<sup>(1)</sup> and Andreas Joseph<sup>(2)</sup>

### Abstract

We introduce machine learning in the context of central banking and policy analyses. Our aim is to give an overview broad enough to allow the reader to place machine learning within the wider range of statistical modelling and computational analyses, and provide an idea of its scope and limitations. We review the underlying technical sources and the nascent literature applying machine learning to economic and policy problems. We present popular modelling approaches, such as artificial neural networks, tree-based models, support vector machines, recommender systems and different clustering techniques. Important concepts like the bias-variance trade-off, optimal model complexity, regularisation and cross-validation are discussed to enrich the econometrics toolbox in their own right. We present three case studies relevant to central bank policy, financial regulation and economic modelling more widely. First, we model the detection of alerts on the balance sheets of financial institutions in the context of banking supervision. Second, we perform a projection exercise for UK CPI inflation on a medium-term horizon of two years. Here, we introduce a simple training-testing framework for time series analyses. Third, we investigate the funding patterns of technology start-ups with the aim to detect potentially disruptive innovators in financial technology. Machine learning models generally outperform traditional modelling approaches in prediction tasks, while open research questions remain with regard to their causal inference properties.

**Key words:** Machine learning, artificial intelligence, big data, econometrics, forecasting, inflation, financial markets, banking supervision, financial technology.

JEL classification: A12, A33, C14, C38, C44, C45, C51, C52, C53, C54, C55, C61, C63, C87, E37, E58, G17, Y20.

<https://www.bankofengland.co.uk/working-paper/2017/machine-learning-at-central-banks>

Irving Fisher Committee  
on Central Bank Statistics

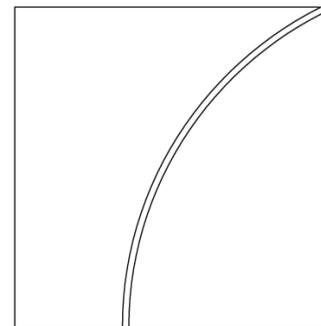
## IFC Bulletin

No 50

## The use of big data analytics and artificial intelligence in central banking

Proceedings of the IFC – Bank Indonesia International Workshop and Seminar in Bali on 23-26 July 2018

May 2019



<https://www.bis.org/ifc/publ/ifcb50.htm>

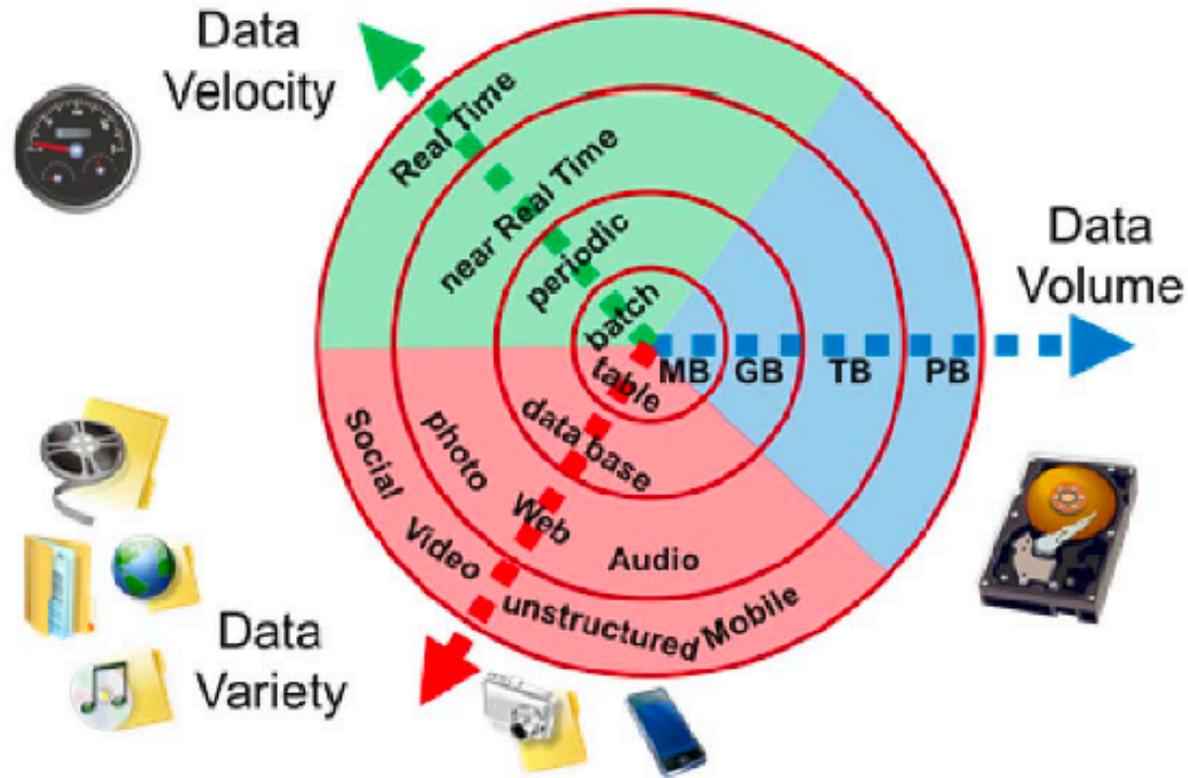
## 2. Why Machine Learning?



# Arguments for Using Machine Learning at Central Banks

- 1. Needed ML Big Data (models with 100+ variables)**
  - “Unstructured” data e.g. satellite imagery, text
- 2. Can learn non-linear relationships**
- 3. Better forecasts / econometrics**
- 4. Anomaly detection (for fraud detection)**

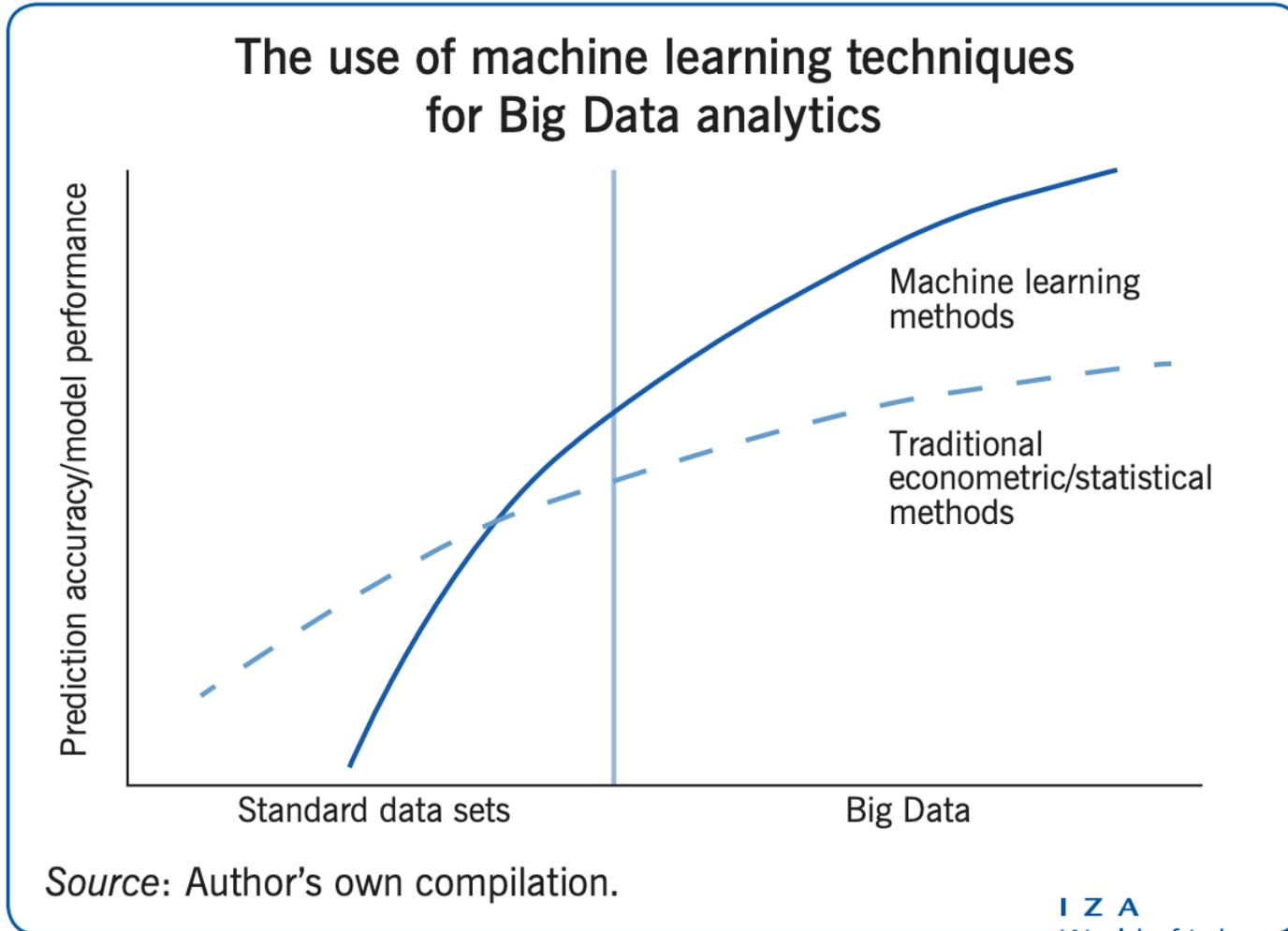
# What is Big Data?



• **Big data is Data with Three “v’s”**

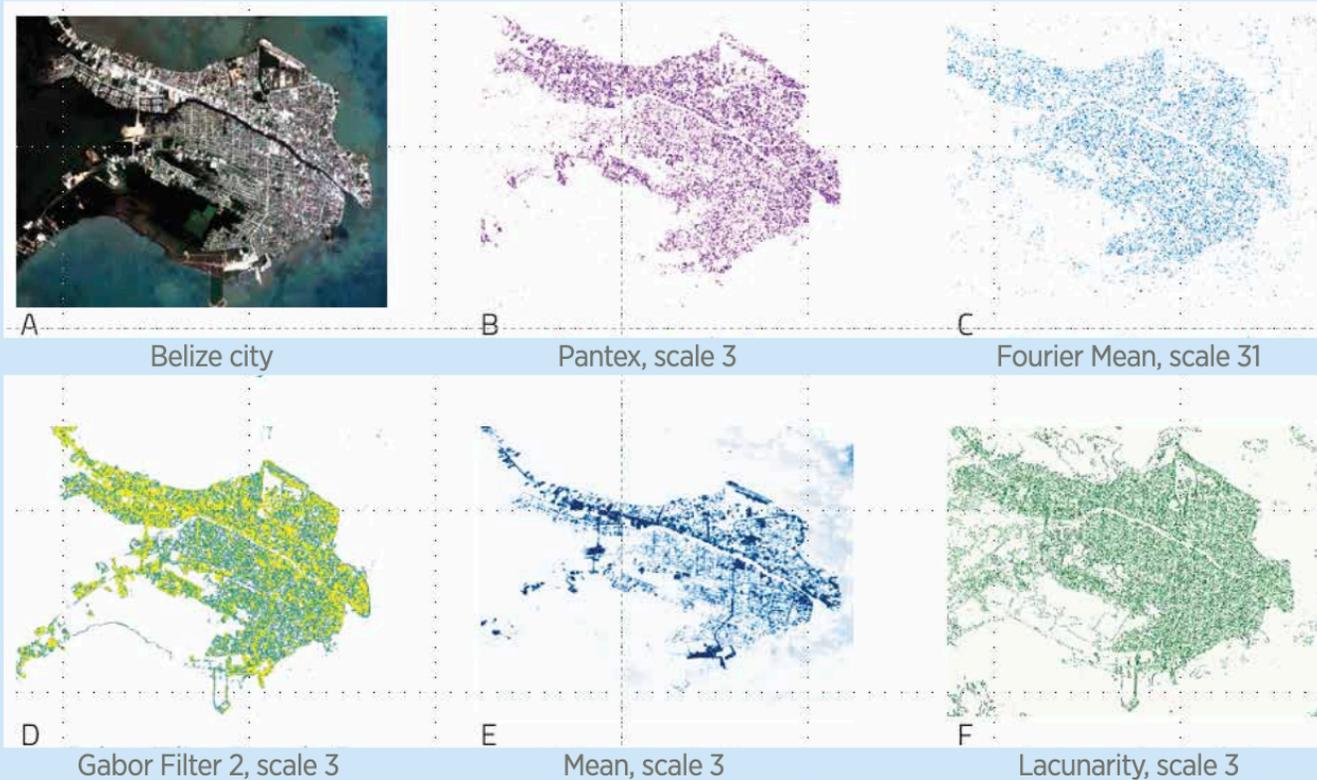
- High volume
- High variety
- High velocity

# Why ML? (Big Data Needs Machine Learning)



- Machine Learning models continue to improve given more data (both # of variables and # of observations)
- Bigger datasets: bigger gain from machine learning vs econometrics

# Why ML? (To Use Satellite Imagery “Big Data”)



**Figure 1** Belize City contextual features

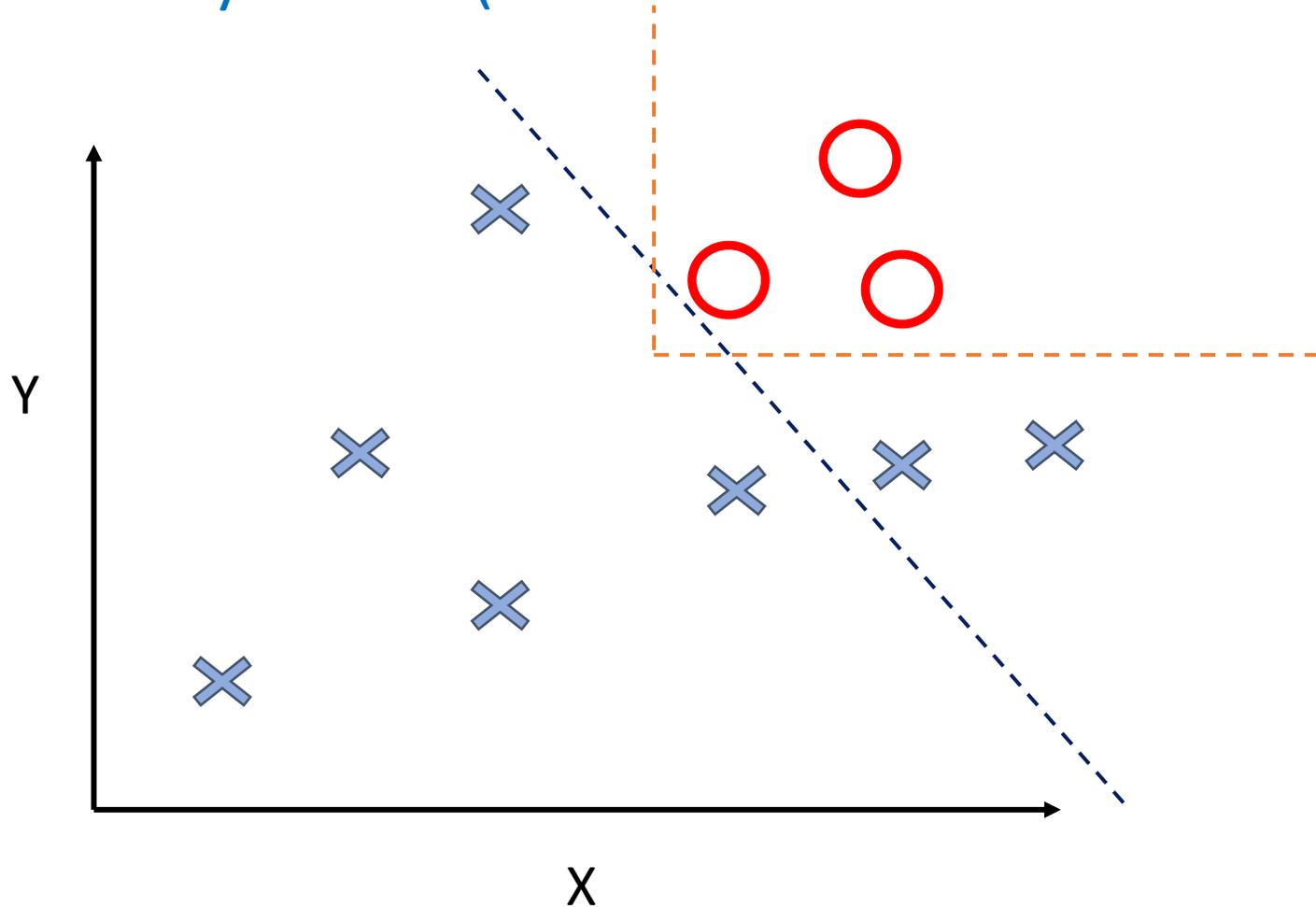
Displayed in **A** is the true color (Blue, Green, and Red) Sentinel 2 image for Belize City, **B** is the Pantex measure derived for this area at scale 3 (30m), **C** is the Mean Fourier transform at scale 31 (310m), **D** is the second Gabor Filter at scale 3 (30m), **E** is the Mean brightness at scale 3 (30m), and **F** is the Lacunarity measure at scale 3 (30m).

- Satellite Imagery variables too high dimensional for traditional econometric models

- Mapping Poverty in Belize Using Satellite Imagery

<https://publications.iadb.org/publications/english/document/Mapping-Income-Poverty-in-Belize-Using-Satellite-Features-and-Machine-Learning.pdf>

# Why ML? (Can learn Nonlinear relationships)



- **Example: classify “O”s separate from X’s**

Econometrics:  $y = X * \beta$

Machine Learning: regression tree

Model	Accuracy
Econometrics	80%
Machine Learning	100%

# Why ML? (Better Forecasts For Fiscal Crises)

## Predicting Fiscal Crises: A Machine Learning Approach

Klaus-Peter Hellwig<sup>1</sup>

International Monetary Fund, Asia Pacific Department

This version: July 27, 2020

**Abstract:** This paper assesses the ability of econometric and machine learning techniques to predict fiscal crises out of sample. We show that the standard econometric approach used in policy applications cannot outperform a heuristic rule of thumb derived from unconditional historical averages. Elastic net and tree ensemble methods (random forest, gradient boosted trees) deliver significant improvements in accuracy. Performance of machine learning techniques improves, particularly for developing countries, when expanding the set of potential predictors from a small set, preselected manually from the literature, to a large set (748 variables) and relying on algorithmic variable selection techniques. There is considerable agreement across learning algorithms in the set of selected predictors: Results confirm the importance of external sector stock and flow variables found in the literature but also point to demographics and the quality of governance as important predictors of fiscal crises. Fiscal variables appear to have less predictive value, and public debt matters only to the extent that it is owed to external creditors.

# Why ML? (Better Forecasts of Inflation)

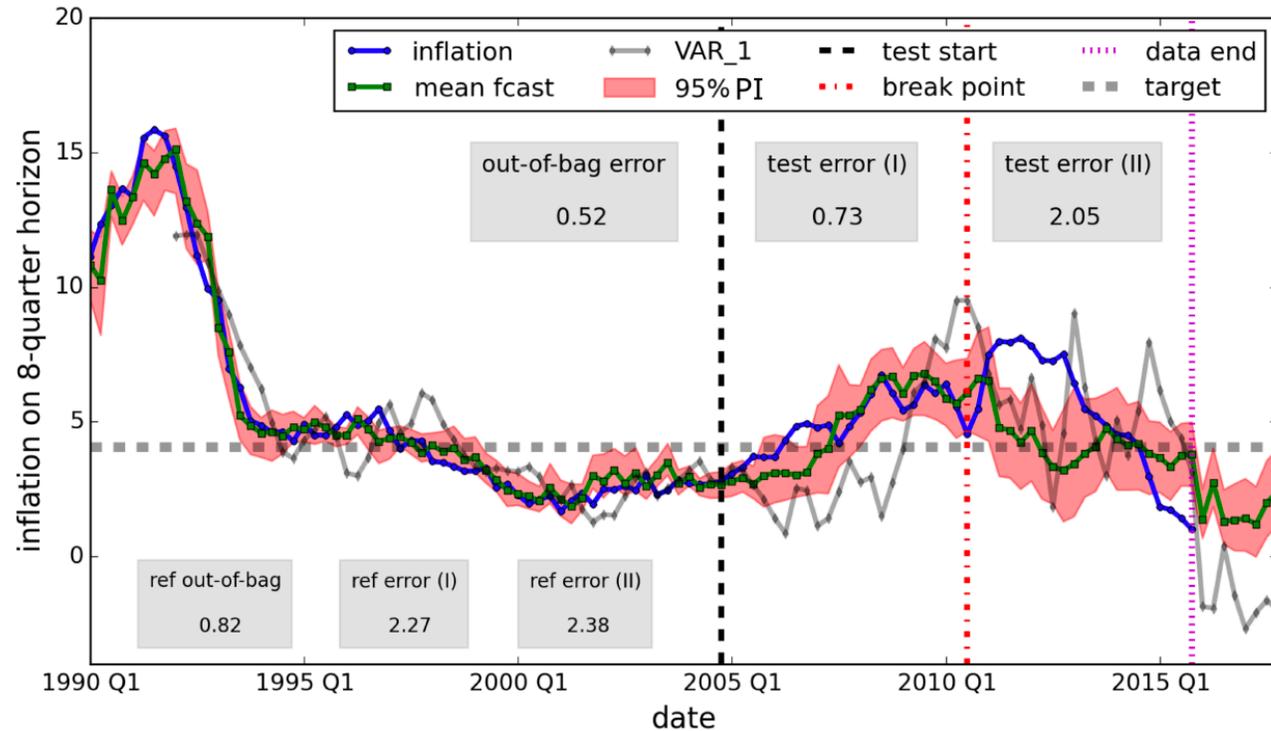
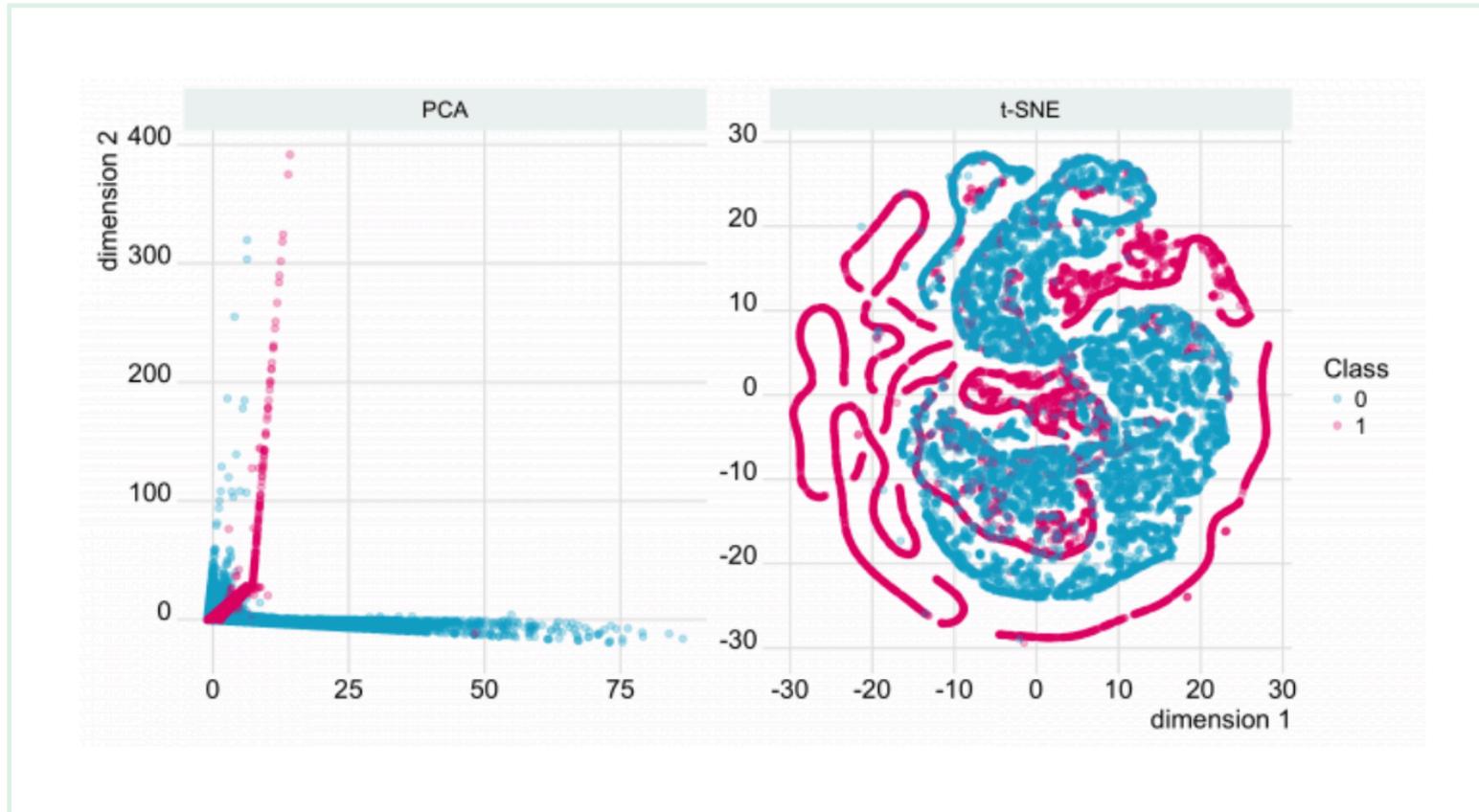


Figure 16: Averaged bootstrapped SVM-dFFANN projections (green line) for two-year changes in CPI (blue line). The shaded band indicates the 95% prediction interval (PI) across bootstrapped models. The vertical dashed lines separates the initial training, pre-crisis, post-crisis and post-data periods, respectively. Model and VAR<sub>1</sub> reference errors are given in the boxes. Sources: BoE, ONS, BIS, World Bank and authors' calculations.

# Why ML? (Anomaly Detection Aka Fraud Alerts)



Dimensionality reduction techniques in fraud analytics. The plots show the first two dimensions of PCA (left) and t-SNE (right) for fraudulent (Class = 1) and regular (Class = 0) transactions.

Source: [https://shiring.github.io/machine\\_learning/2017/05/01/fraud](https://shiring.github.io/machine_learning/2017/05/01/fraud)

# 3. Some Basic Machine Learning Concepts



# Supervised vs Unsupervised Learning

## Supervised Learning:

- For every  $x_i$  we observe some  $y_i$
- Ex: random forests to predict loan default ( $y_i$ ) based on applicant characteristics ( $x_i$ )

Supervised Learning



Unsupervised Learning



## Unsupervised Learning:

- We only observe  $x_i$
- Ex: clustering loan applicants based on characteristics ( $x_i$ )

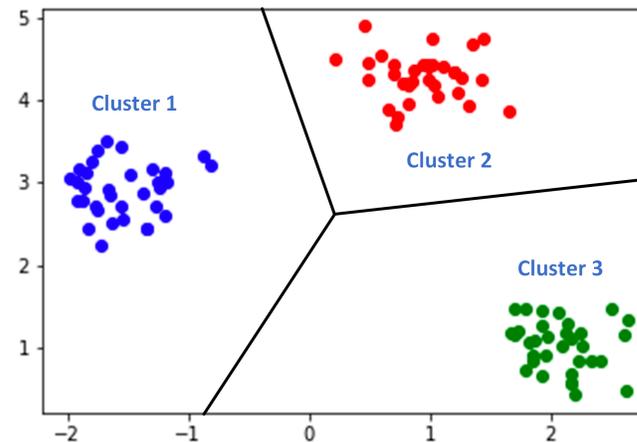
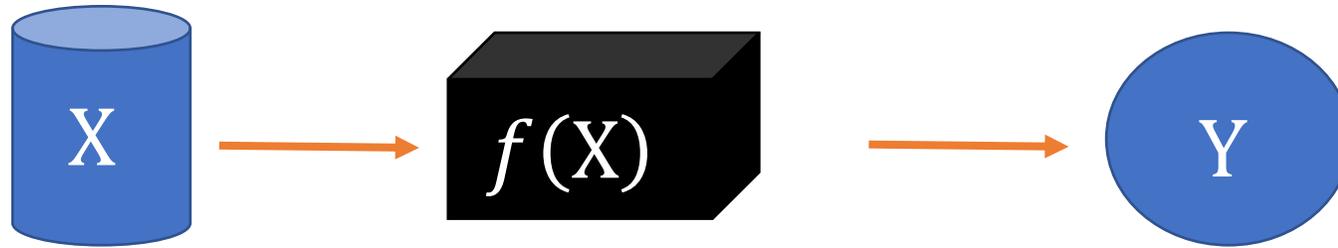


Fig.1. An Example Of Data Clustering

Supervised learning: learning  $f(X)$  our predicted out given inputs

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



$\epsilon$  = “epsilon” (unexplained portion)

# “Estimating” $\hat{f}(X)$

- $Y = f(X) + \epsilon$  is the true value
- We can only use data to “guess” at  $f(X)$
- We call this guess  $\hat{f}(X)$

## **How do we know when we’ve selected a “good” $\hat{f}(X)$ ?**

- We reserve a portion of our data into a “test” set, estimate a model on the other part, and see how our model performs on this test set

# Testing Training Data Subsets

**Training set:** (observation-wise) subset of data used to develop models



# Testing/Training Split

**Training set:** (observation-wise) subset of data used to develop models

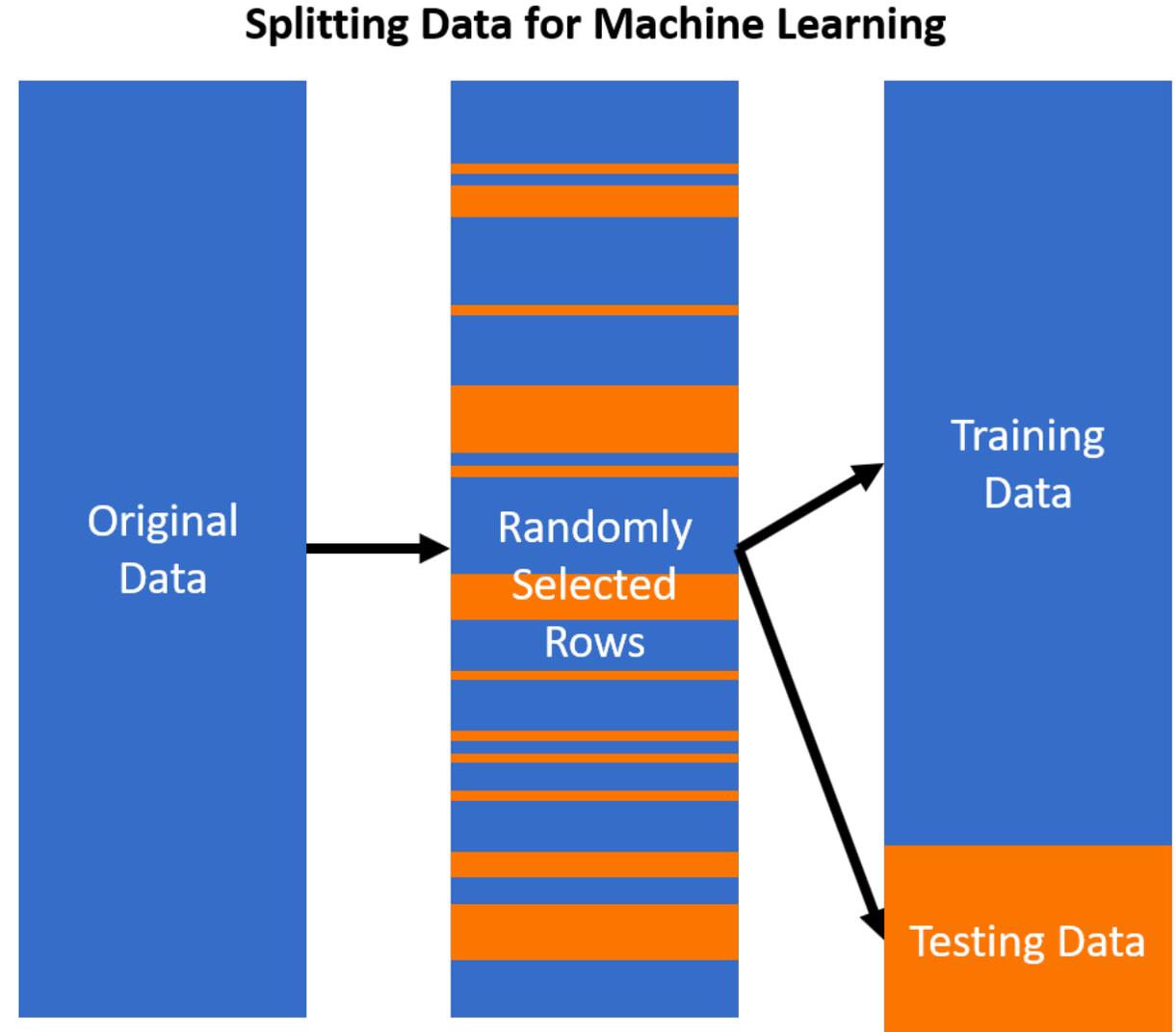
**Test set:** subset of data used during intermediate stages to “tune” model parameters

**Rule of thumb 75% training 25% test -ish**



# Randomly Selecting Rows for Test or Training Sets

- Observations are randomly selected into either testing or training splits of the data



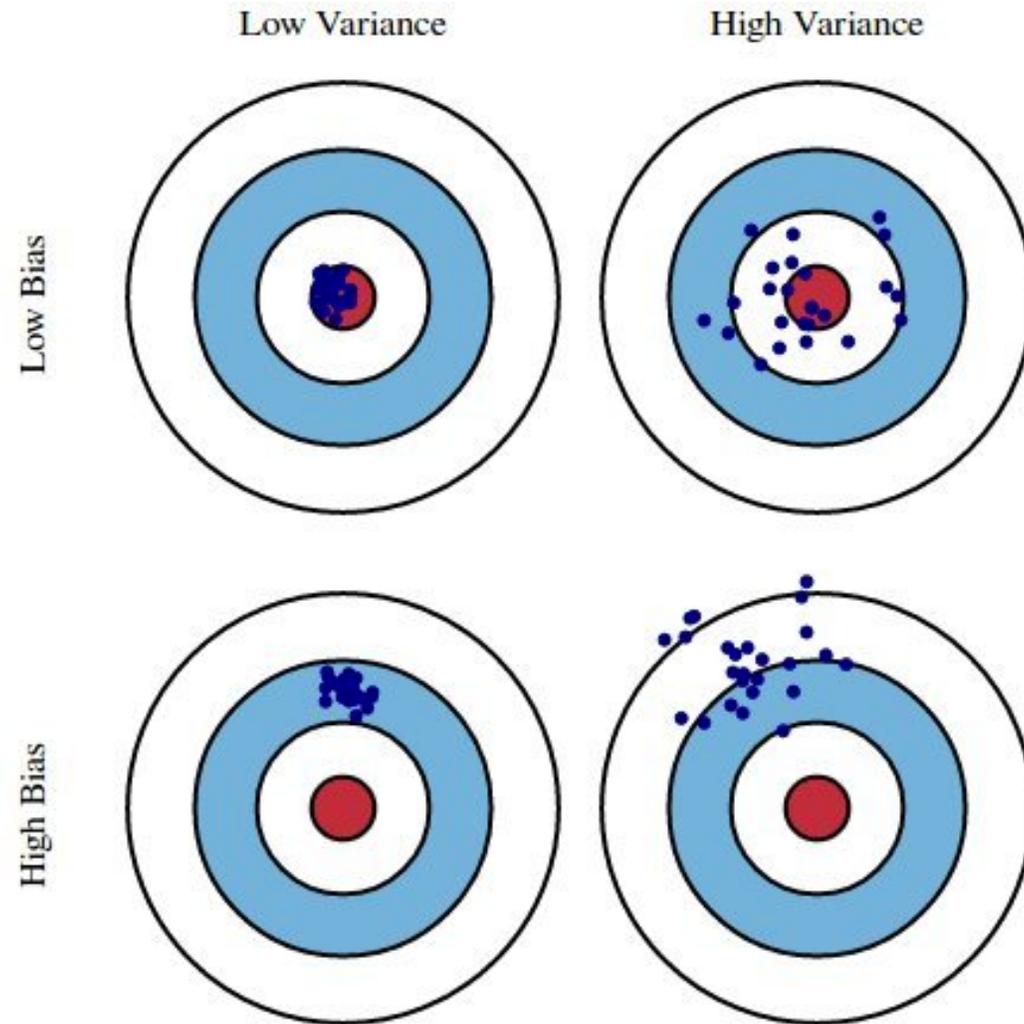
# Bias and Variance

**Bias: Tendency of an in-sample statistic to over or under estimate the statistic in the *population***

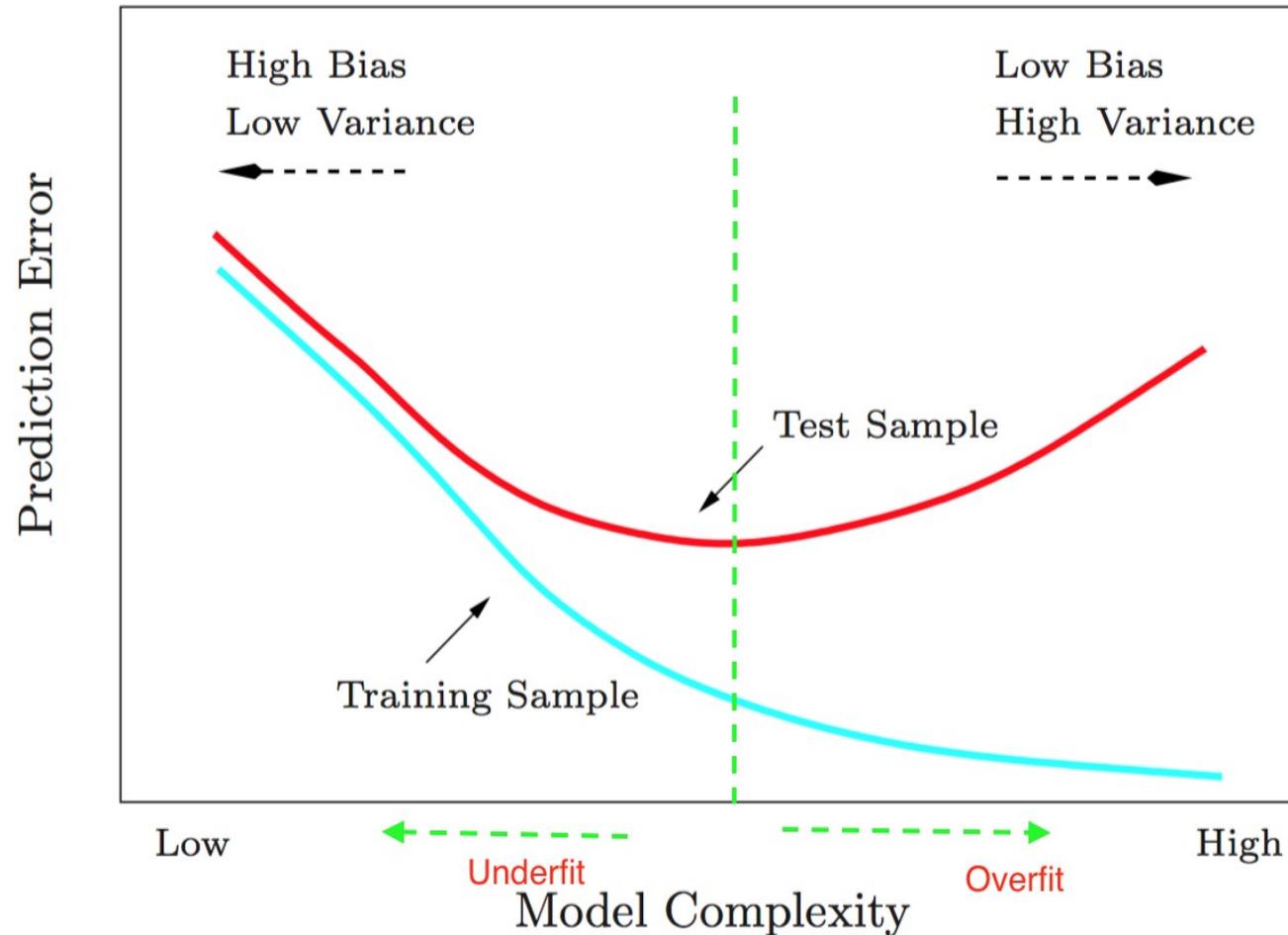
**Variance: Tendency to noisily estimate a statistic.**

E.g., sensitivity to small fluctuations in the training dataset.

# Bias-Variance Tradeoff

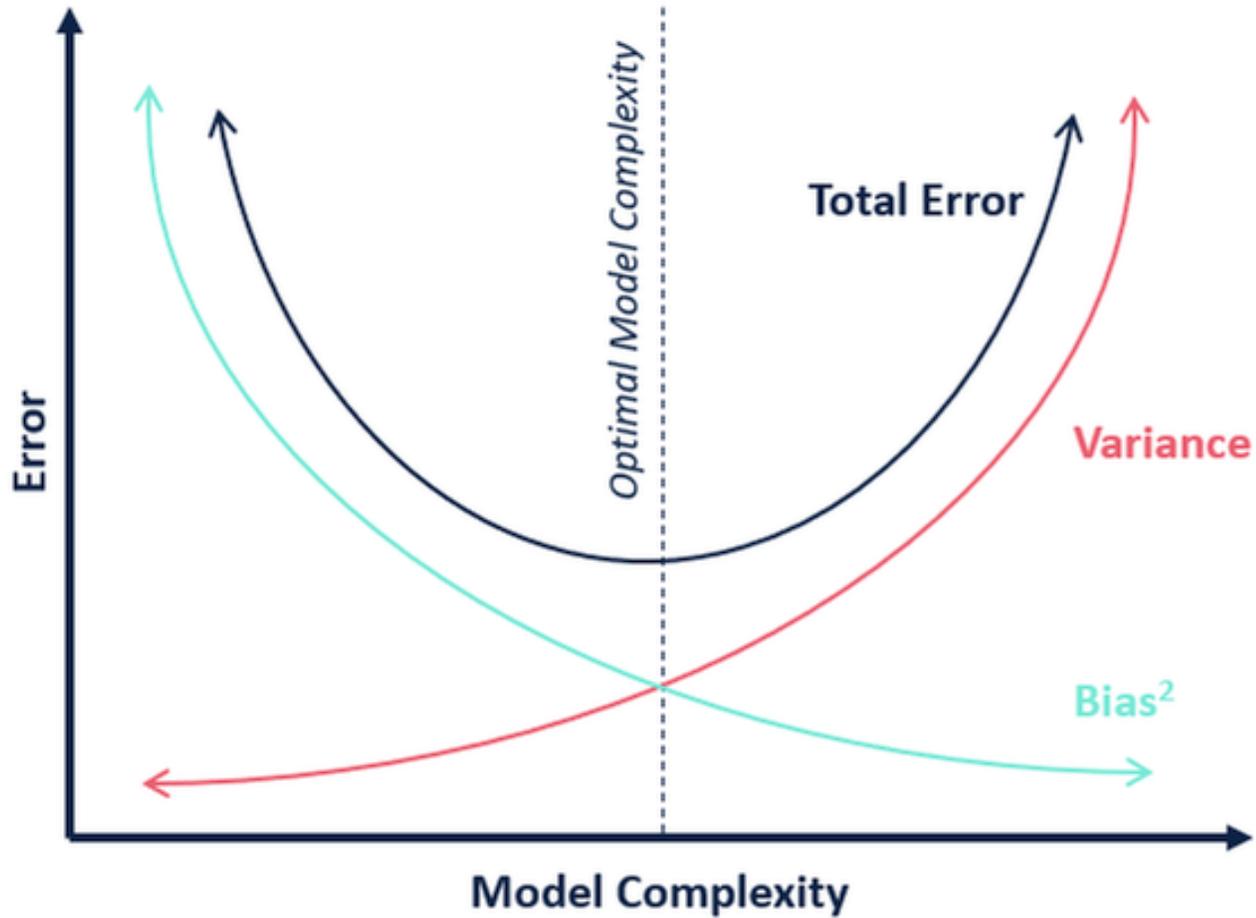


# Bias-Variance Tradeoff

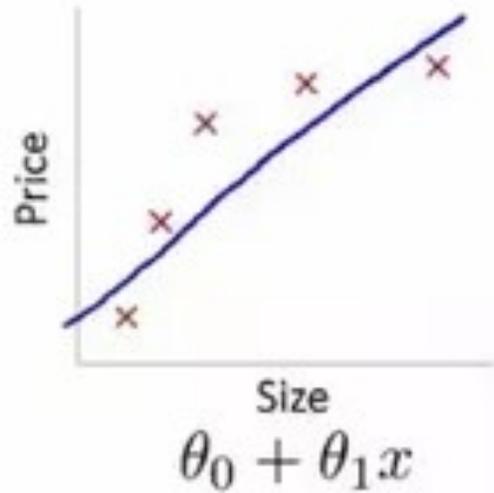


- Error in Training sample ( $\sim$ bias)  $\downarrow$  as we  $\uparrow$  model complexity (e.g. number of variables)
- Error in Test sample ( $\sim$ variance)  $\uparrow$  as we  $\uparrow$
- Key: finding optimal model complexity

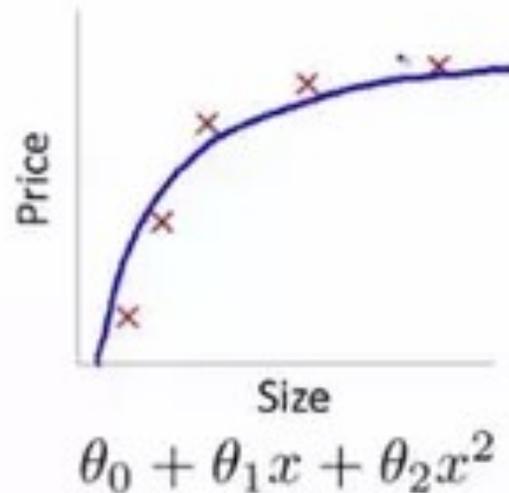
# Key: Finding Optimal Model Complexity



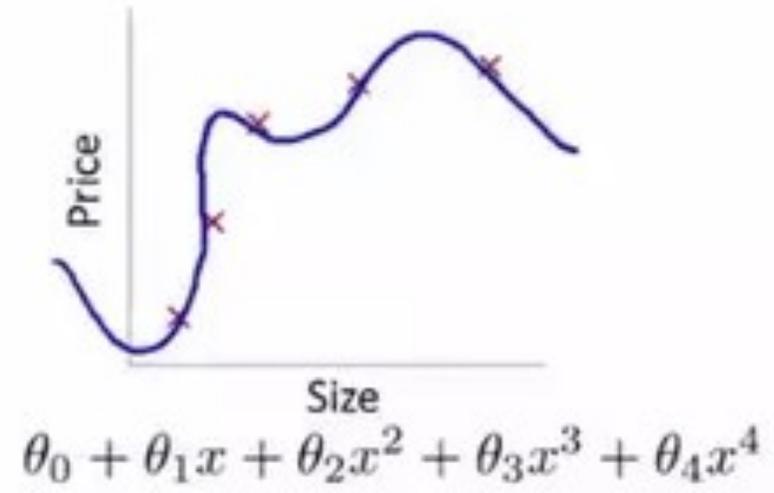
# Optimal Model Complexity: Neither Underfit Nor Overfit



High bias  
(underfit)



“Just right”



High variance  
(overfit)

# Assessing Model Accuracy: Mean Squared Error

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

$\sum$  means we add up anything with  $i$ , starting at  $i = 1$  to  $i = n$

$y_i$	$\hat{y}_i$	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$
5	5	0	$0^2 = 0$
6	7	-1	$-1^2 = 1$
9	8	1	$1^2 = 1$
10	1	9	$9^2 = 81$

# Summary – Intro to Machine Learning

- **Machine Learning** is a set of methods developed to find robust patterns across datasets
- **Central Banks can benefit from machine learning.**
  - Big data requires it
  - Non-linear relationships
  - Better forecasts/econometrics
  - Anomaly detection
- **Remember these key concepts**
  - Supervised (Y,X) vs Unsupervised learning (just X)
  - Testing/Training Sets
    - (model -> train, see how it performs on test)
  - Bias-Variance Tradeoff
    - Bias – how far off model from true
    - Variance – precision of estimated model
    - Total error =  $\text{bias}^2 + \text{variance}$