3. Introduction to Machine Learning for Public Policy

Jonathan Hersh, PhD (Chapman Argyros School of Business)

Outline

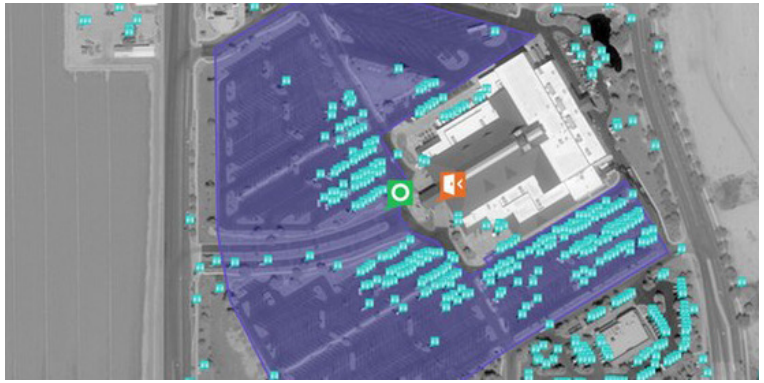
1. **What is Machine Learning?**
   - Machine learning versus econometrics
2. **Why Machine Learning for Public Policy**
   - Big data requires it
   - Non-linear relationships
   - Better forecasts/econometrics
   - Anomaly detection
3. **Some Basic Machine Learning Concepts**
   - Supervised vs Unsupervised learning
   - Testing/Training Sets
   - Bias-Variance Tradeoff

About Me

- Assistant Professor Economics and Management Science Chapman University
- PhD in economics, Boston University
- **Research Fields**:
  - Applications of artificial intelligence (computer vision)
  - Economics of information systems
  - Development economics
  - Digitization strategy
- **Teaching Fields**:
  - Machine learning
  - Applications of artificial intelligence

My Research

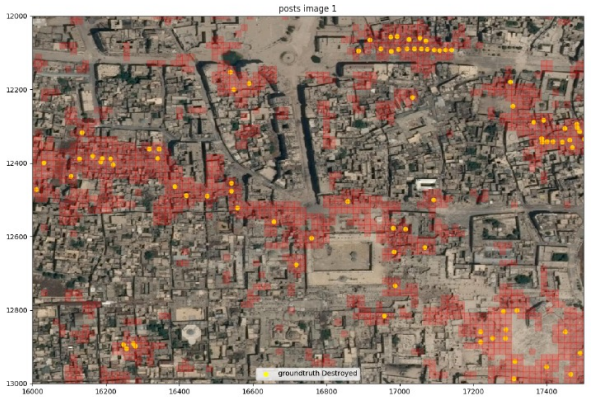
- Satellite Imagery + Computer Vision + Machine Learning



Count cars in parking lots!

Dense Prediction: Scanning Aleppo

Damaged buildings in Syria!



- Advised World Bank/IDB on COVID poverty transfers in Belize, Togo, Guinea

FAST COMPANY

COVID-19 CO.DESIGN TECH WORK LIFE CREATIVITY IMPACT PODCASTS VIDEO RECOMME

11-06-15 | ELASTICITY

How Satellite Data And Artificial Intelligence Could Help Us Understand Poverty Better

New technology lets computers understand what they see in an image–or a million im

[PHOTO: FLICKR USER RODRIGO CARVALHO]

BY MAYA CRAIG 3 MINUTE READ

Data analytics firm Orbital Insight is partnering with the World Bank to test technology that could help measure global poverty using satellite imagery and artificial intelligence.

Bloomberg

Economics

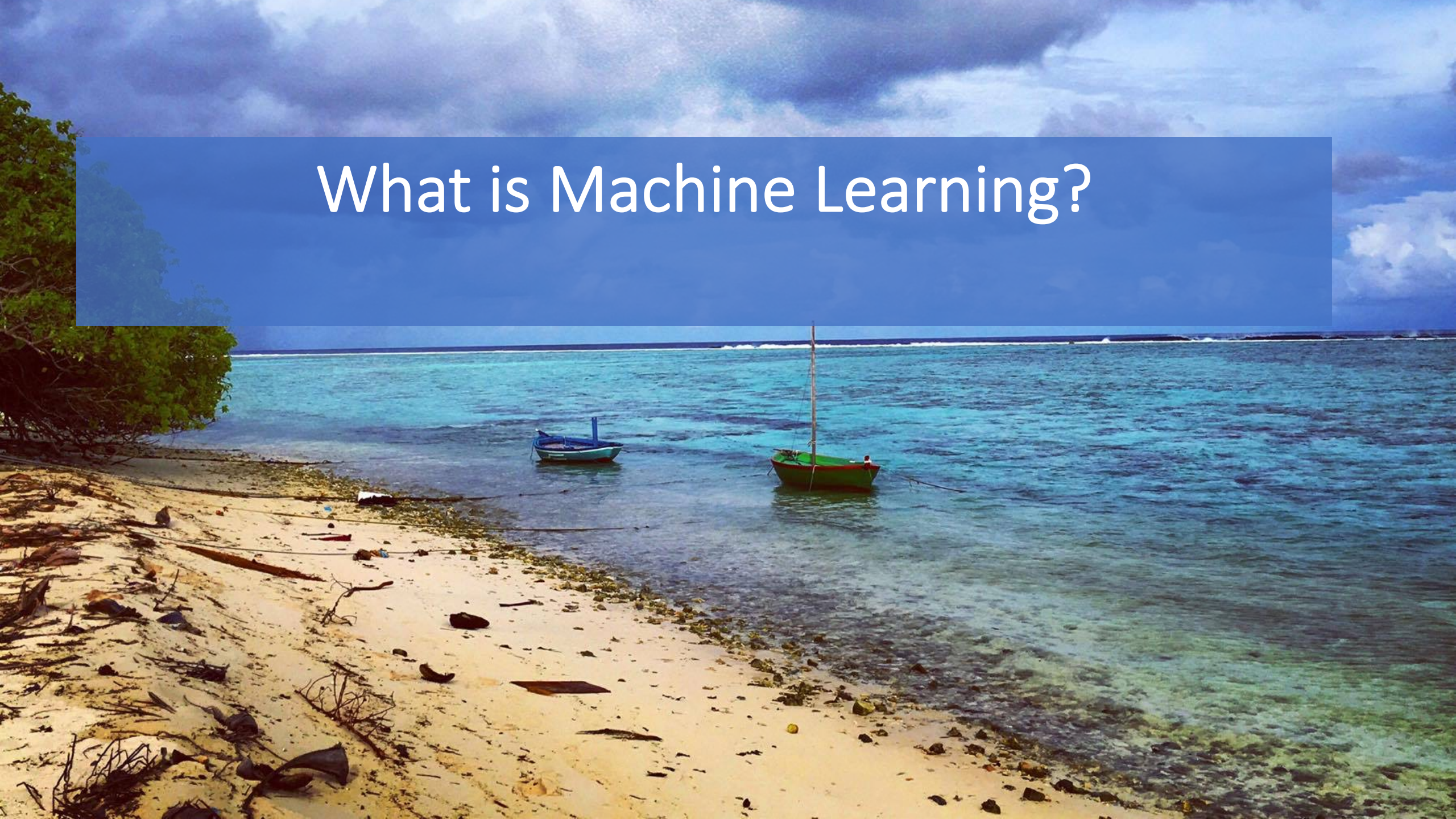
Poverty Surveyors in Sri Lanka Get Some Help From Satellites Orbiting the Earth

The World Bank is teaming with a Silicon Valley startup to test whether poverty can be measured using satellite images.

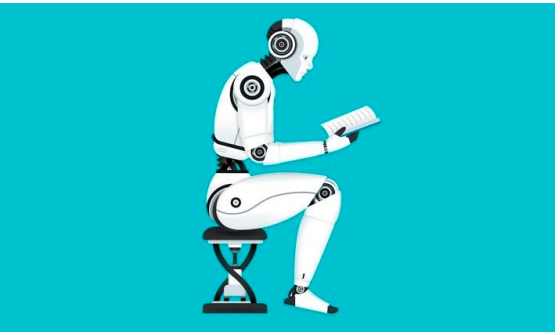
By Adam Satariano

November 6, 2015, 7:00 AM PST *Updated on November 6, 2015, 1:57 PM PST*

In mountainous areas of Pakistan or far-flung villages in Sri Lanka, finding reliable economic information is extremely difficult. The World Bank's solution has been to send surveyors to study the conditions on the ground, which is an expensive, time-consuming, and imprecise task. The resulting dearth of data leaves governments, aid groups, and researchers unsure of where to put resources that can be critical to helping the world's most impoverished areas.


What is Machine Learning?

Public Conception of Machine Learning

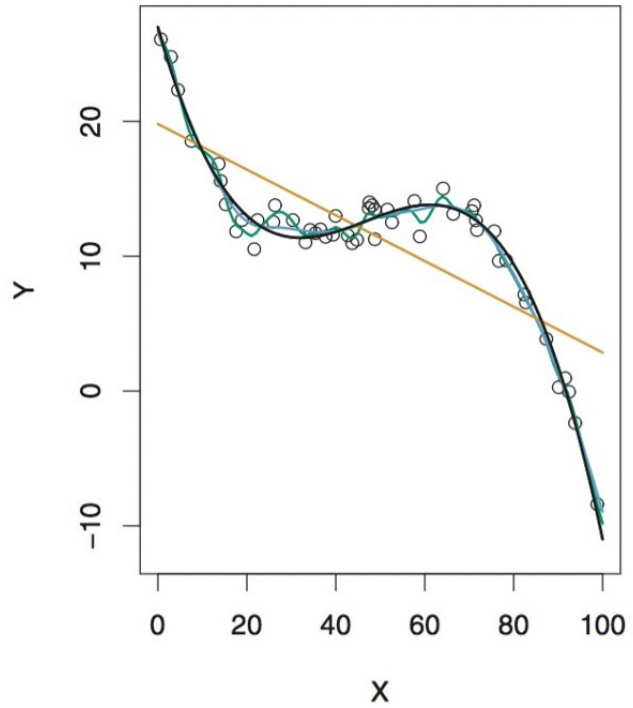


Reality (90% of the time)

Target or Output

Input data

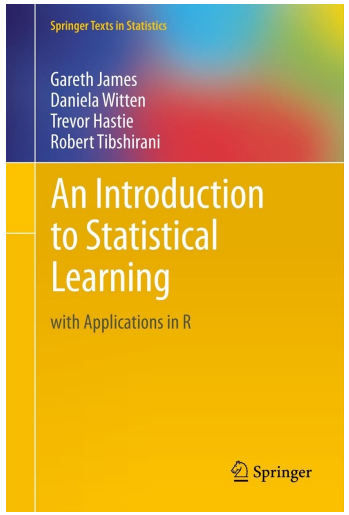
$$\hat{y} = \hat{f}(x)$$

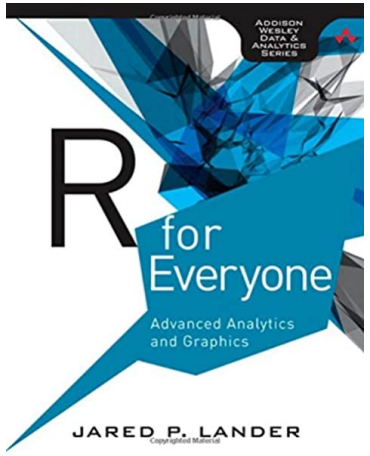


Machine Learning Versus Econometrics

- **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?”

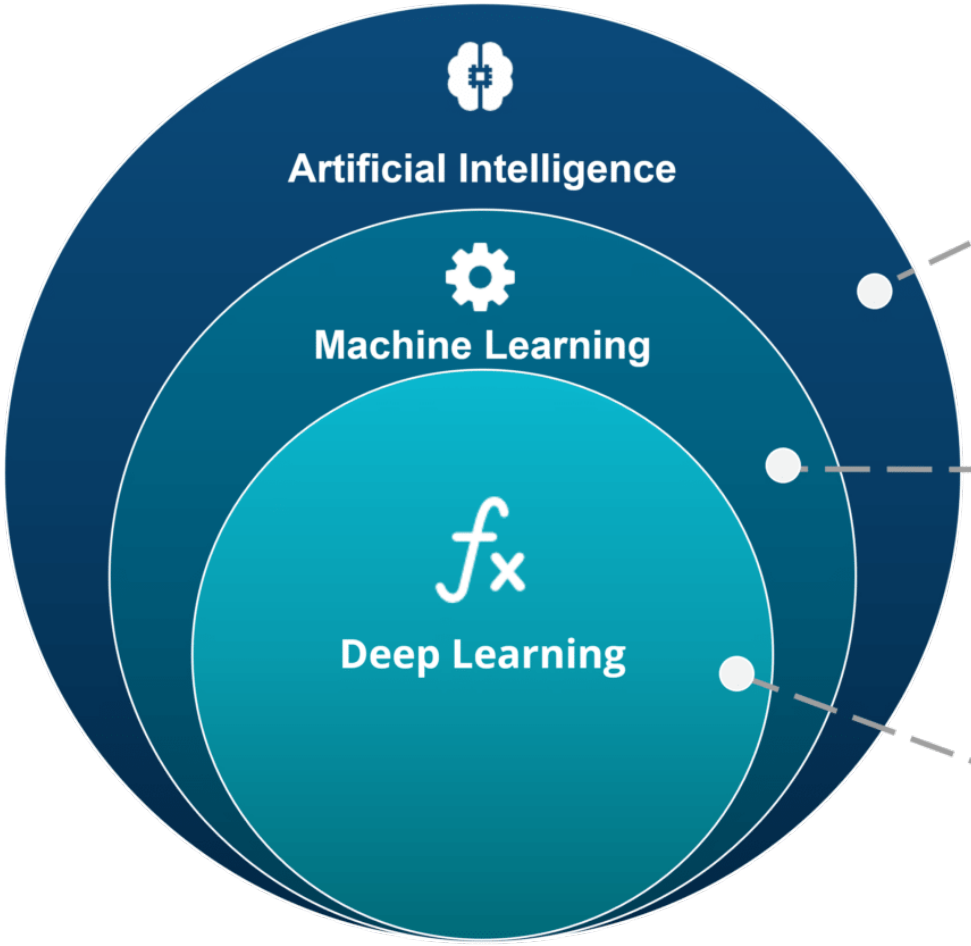
Today – Brief Introduction to Machine Learning





- **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 Everyone]

Machine Learning Versus Artificial Intelligence



ARTIFICIAL INTELLIGENCE
A technique which enables machines to mimic human behaviour

MACHINE LEARNING
Subset of AI technique which use statistical methods to enable machines to improve with experience

DEEP LEARNING
Subset of ML which make the computation of multi-layer neural network feasible


Why Machine Learning?

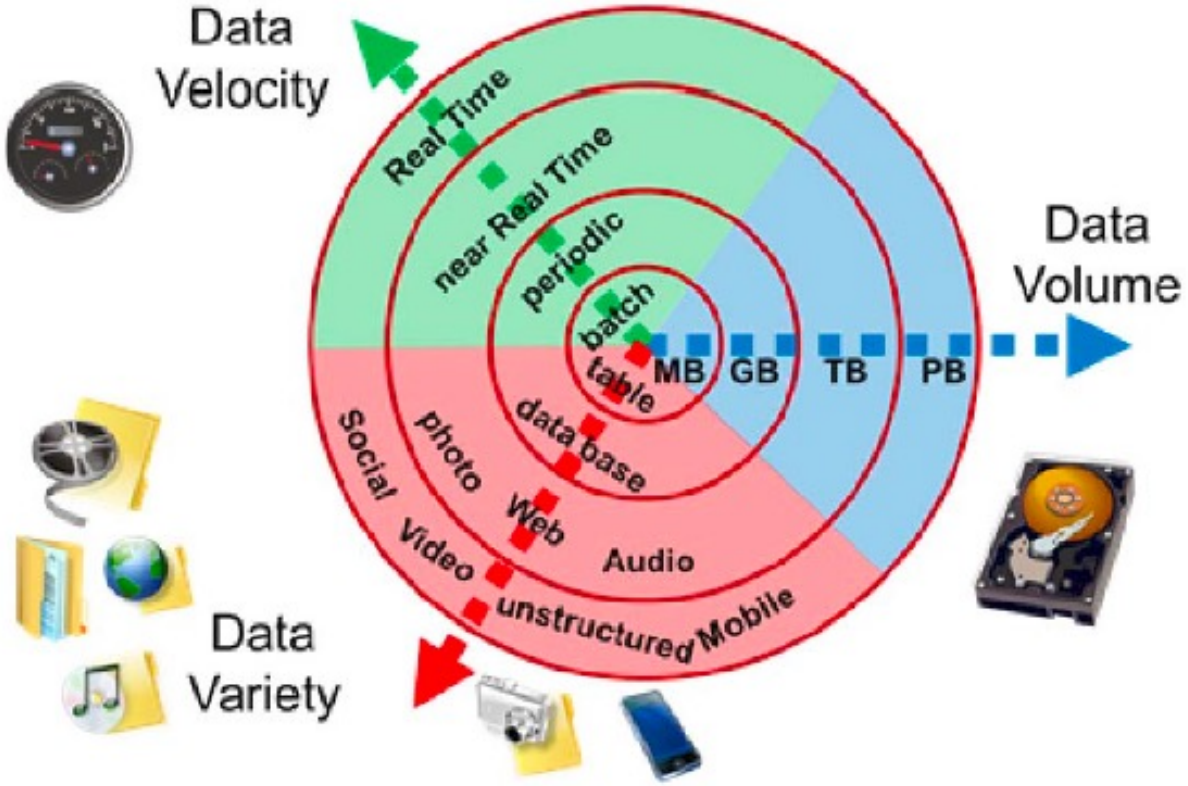
Arguments for Using Machine Learning for Public Policy

1. **Needed ML Big Data (models with 100+ variables)**
   - "Unstructured" data e.g. satellite imagery, text

2. **Can learn non-linear relationships**

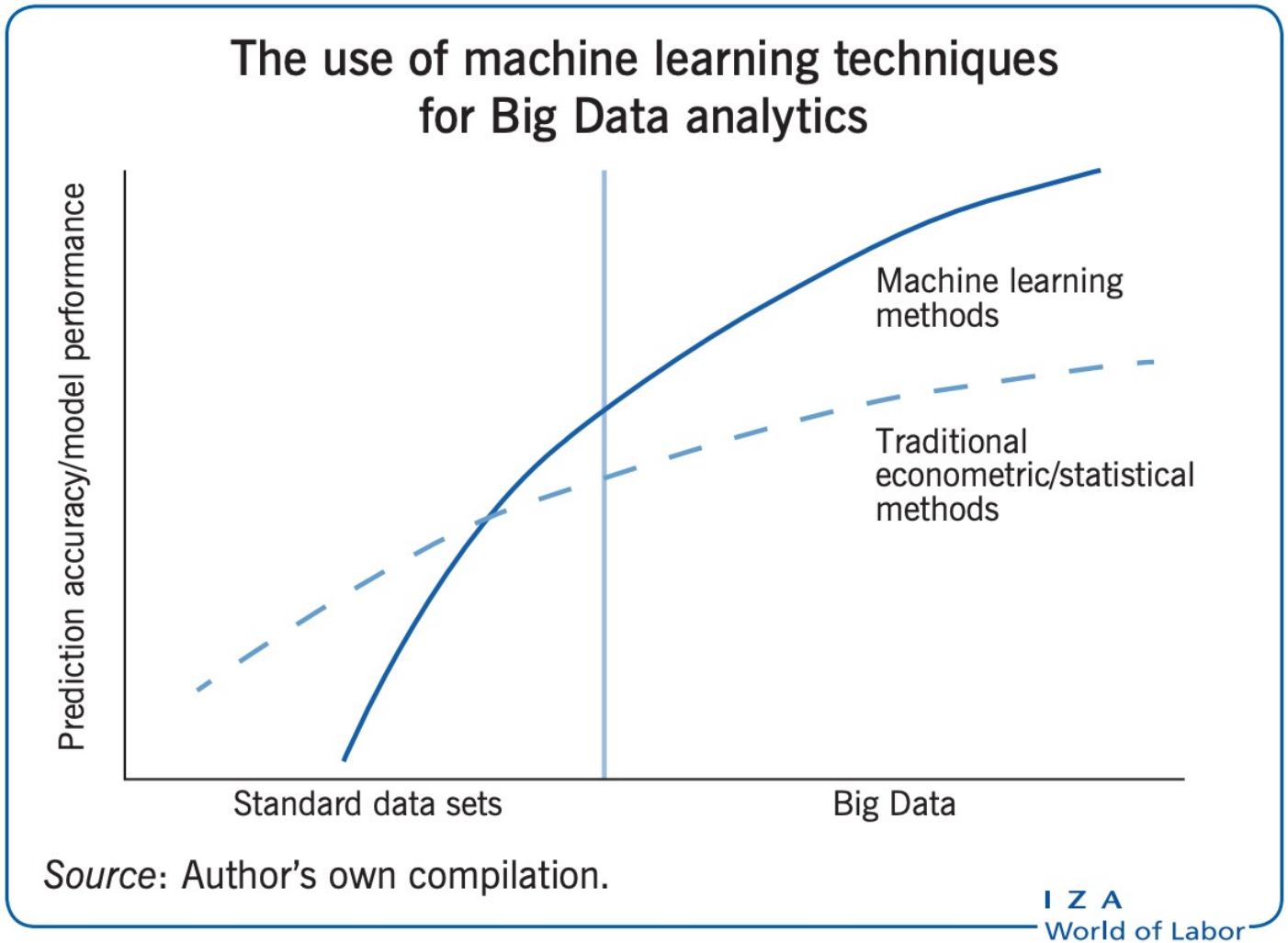
3. **Better forecasts / econometrics**

What is Big Data?



- **Big data is Data with Three "v's"**
  - High <u>volume</u>
  - High <u>variety</u>
  - High <u>velocity</u>

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”)

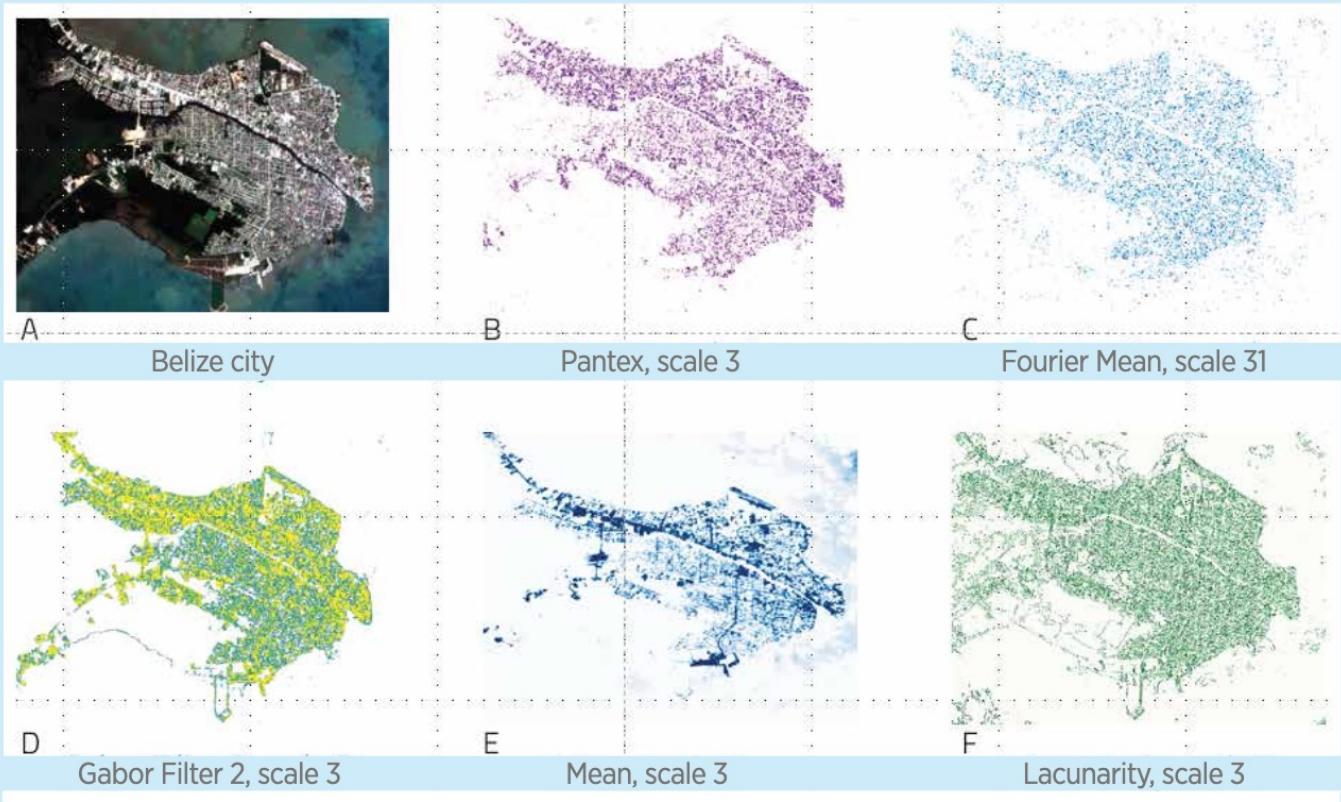
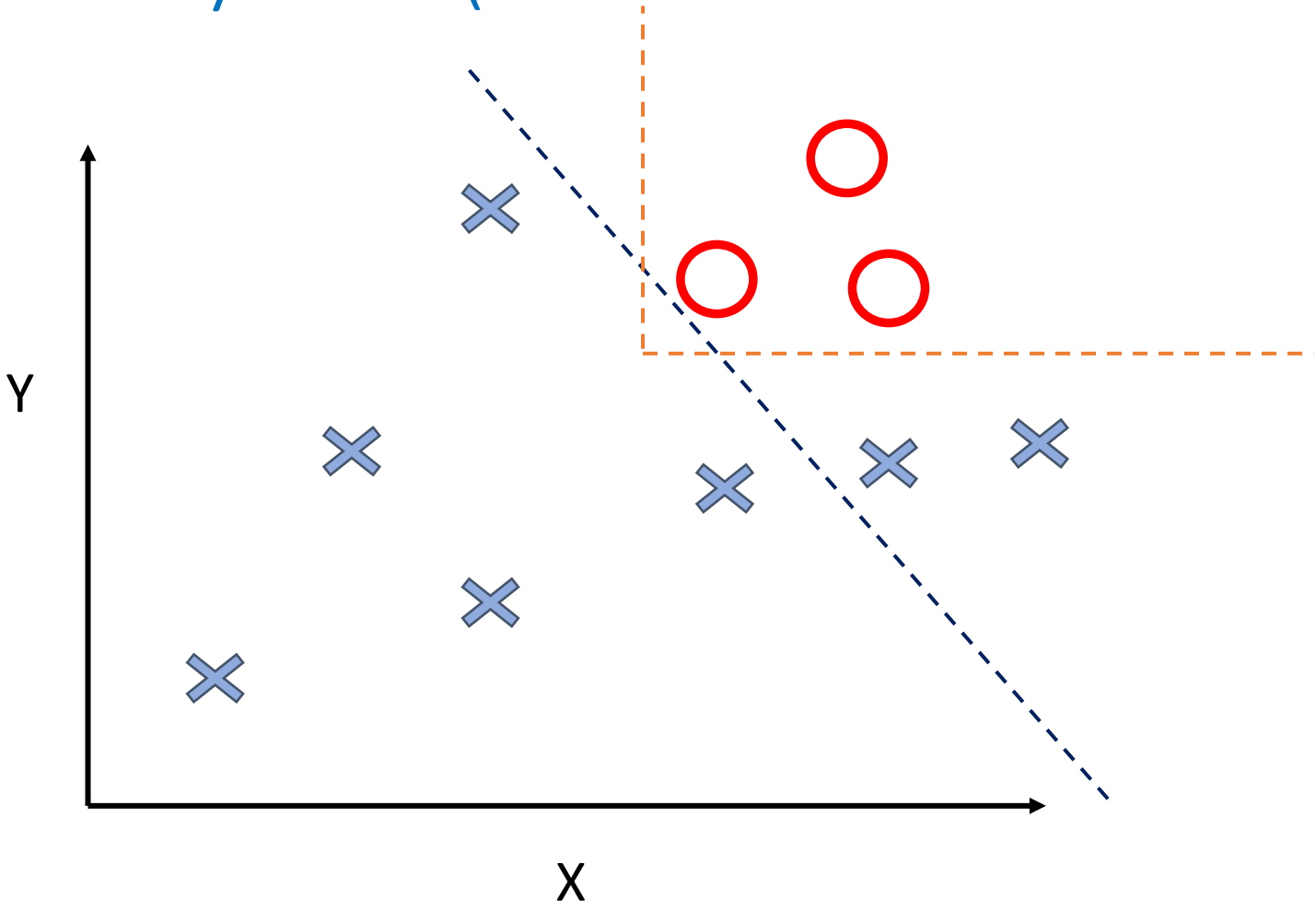


Figure 1 *Belize City contextual features*
*Displayed in **A** is the true color (Blue, Green, and Red) Sentinel 2 image for Belize City, **B** is the Pantex measure derived for this area at scale 3 (30m), **C** is the Mean Fourier transform at scale 31 (310m), D) is the second Gabor Filter at scale 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)



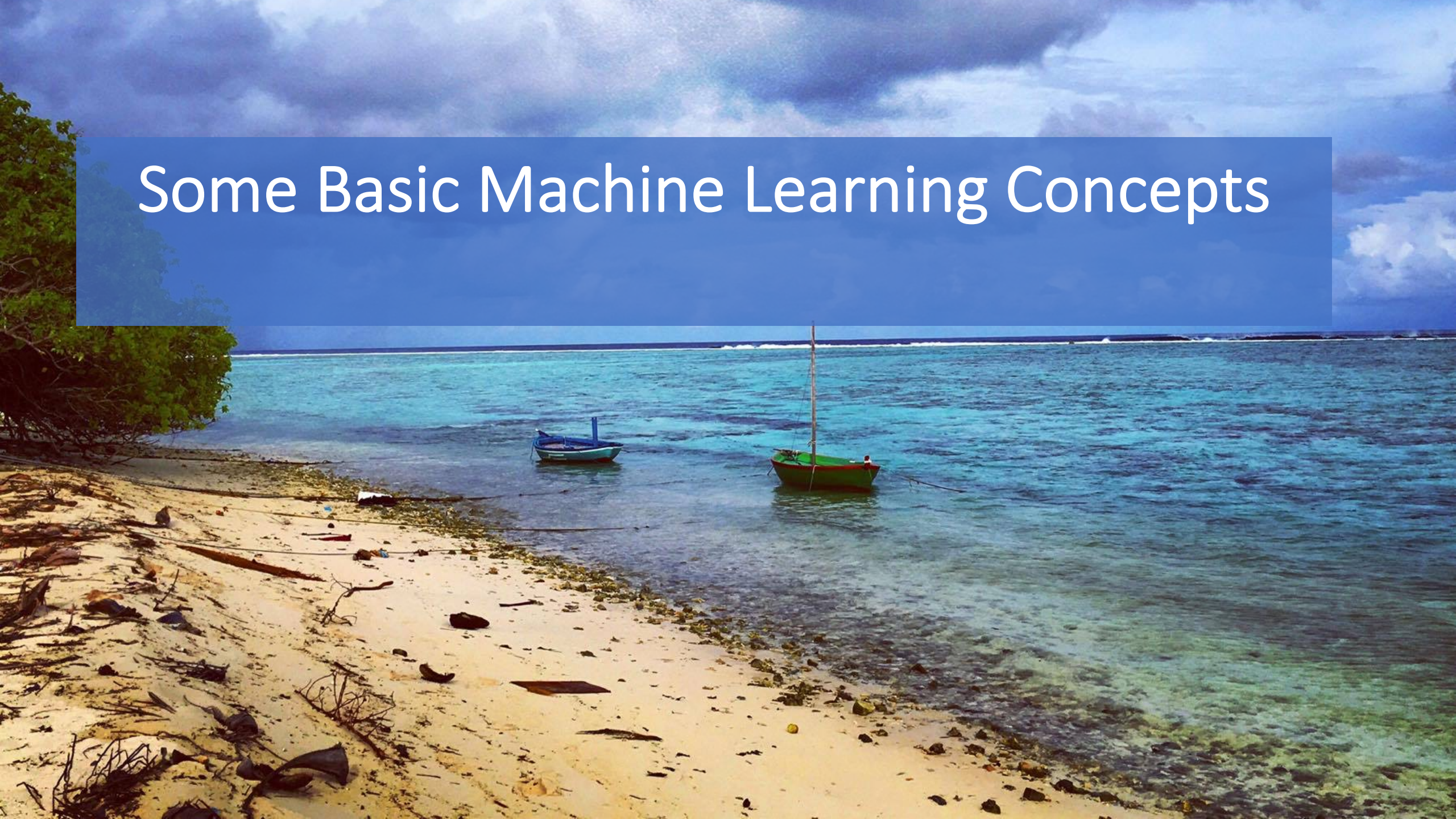
- **Example: classify "O"s separate from X's**

Econometrics: $y = X * \beta$

Machine Learning: regression tree

Model	Accuracy
Econometrics	80%
Machine Learning	100%

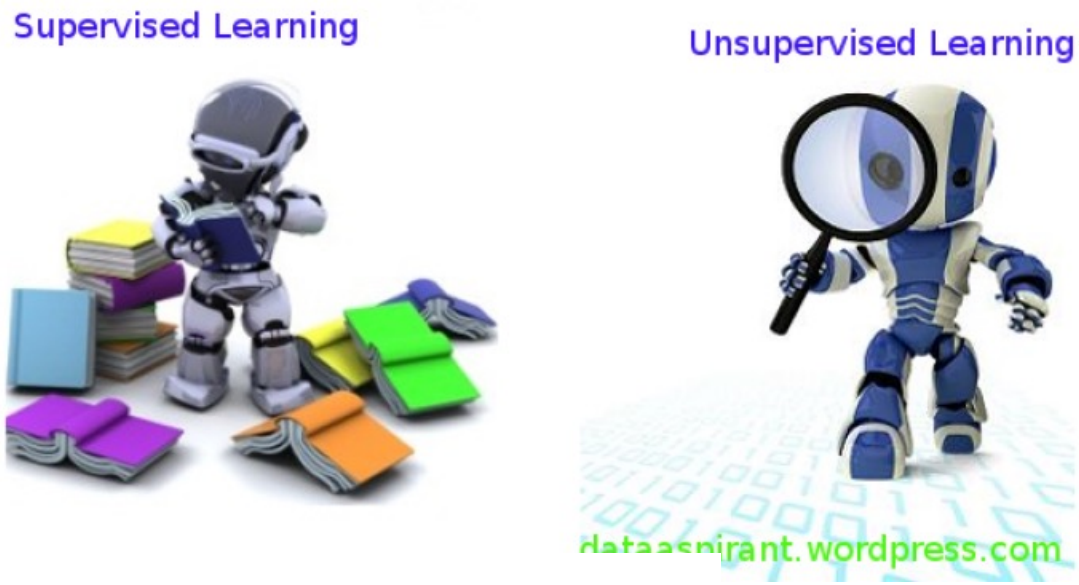
Some Basic Machine Learning Concepts



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)



Unsupervised Learning:

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

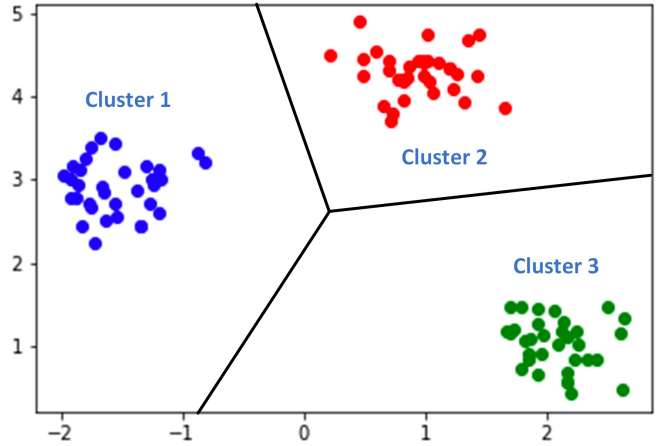
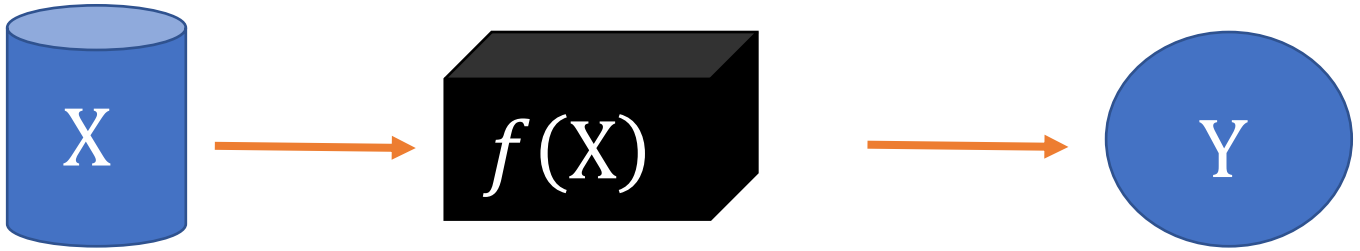


Fig.1. An Example Of Data Clustering

Supervised learning: learning $f(\mathrm{X})$ our predicted out given inputs

$$\mathrm{Y} = f(\mathrm{X}) + \epsilon$$



ϵ = “epsilon” (unexplained portion)

"Estimating" $\hat{f}(X)$

- $\mathrm{Y} = f(\mathrm{X}) + \epsilon$ is the true value
- We can only use data to "guess" at $f(\mathrm{X})$
- 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

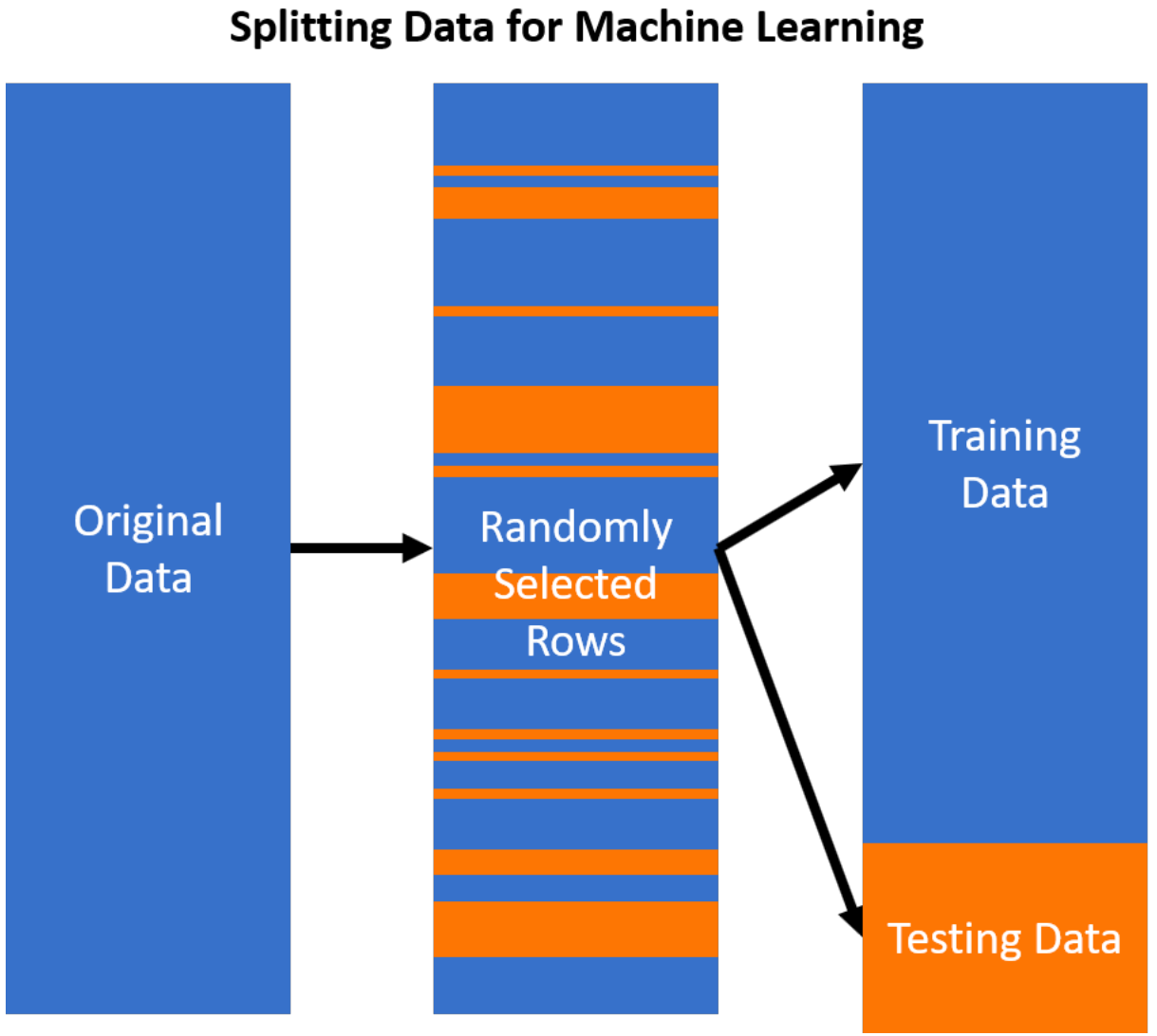
Training set: (observation-wise) subset of data used to develop models

Test or Validation 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



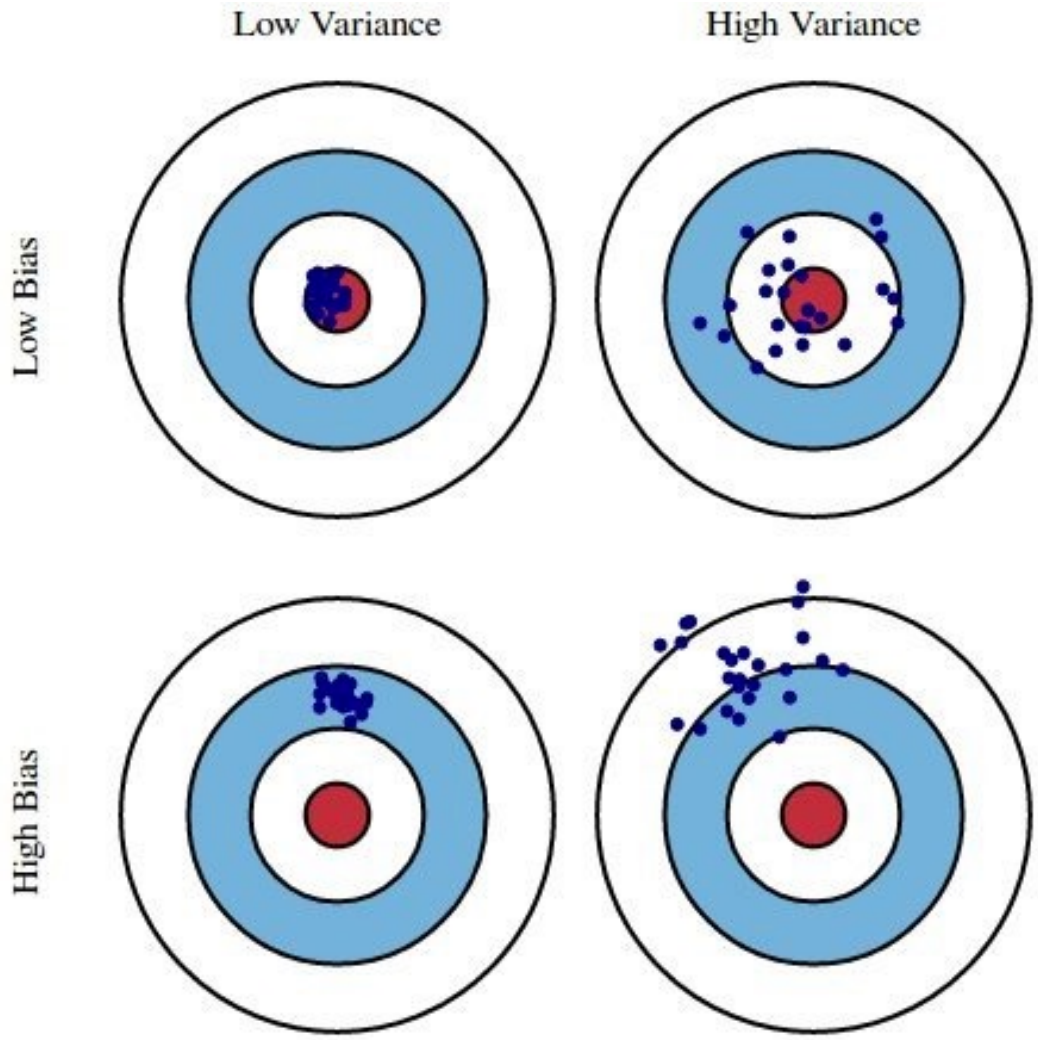
Bias and Variance

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

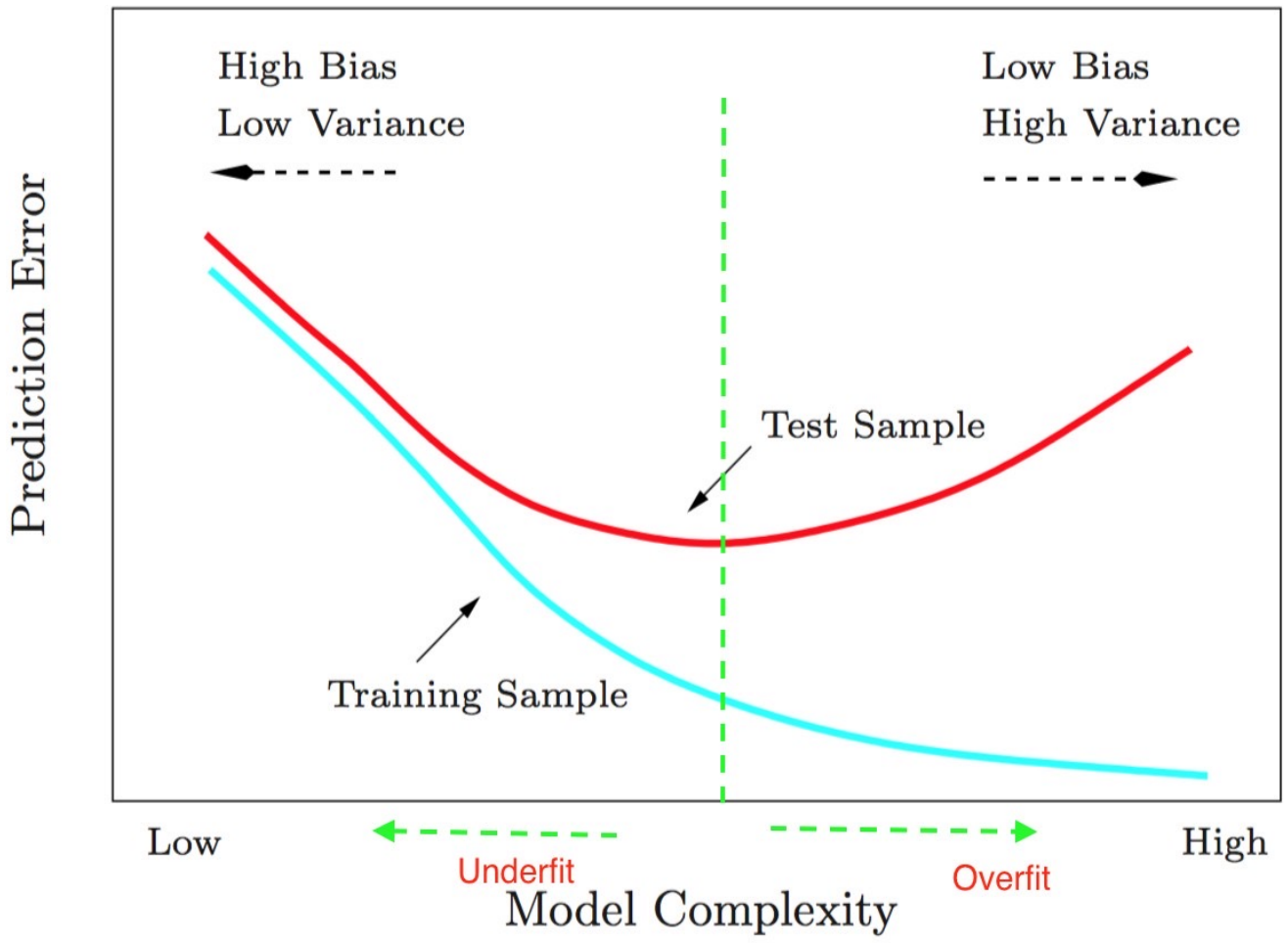
Variance: **Tendency to noisily estimate a statistic**.

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

Bias-Variance Tradeoff

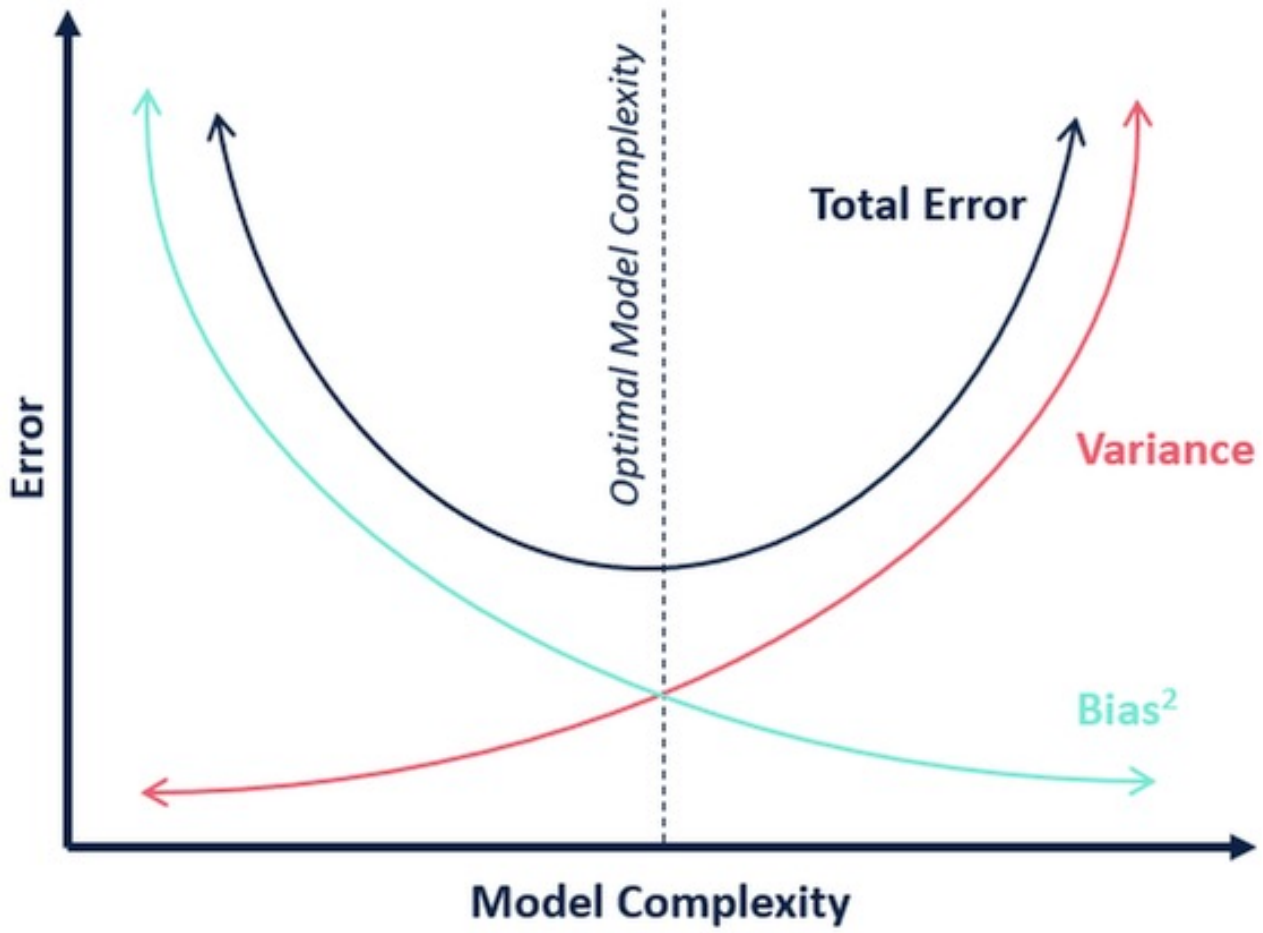


Bias-Variance Tradeoff

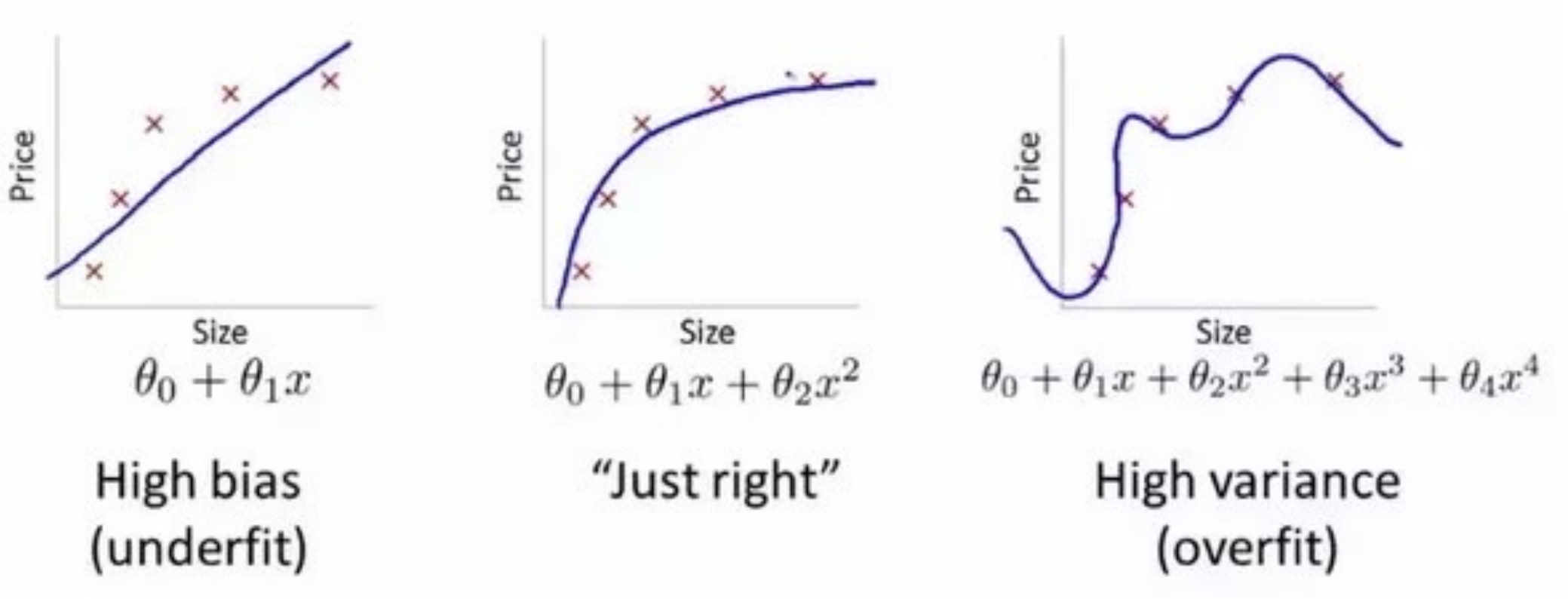


- Error in Training sample (~bias) ↓ as we ↑ model complexity (e.g. number of variables)
- Error in Test sample (~variance) ↑ as we ↑
- Key: finding optimal model complexity

Key: Finding Optimal Model Complexity



Optimal Model Complexity: Neither Underfit Nor 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 $i = n$

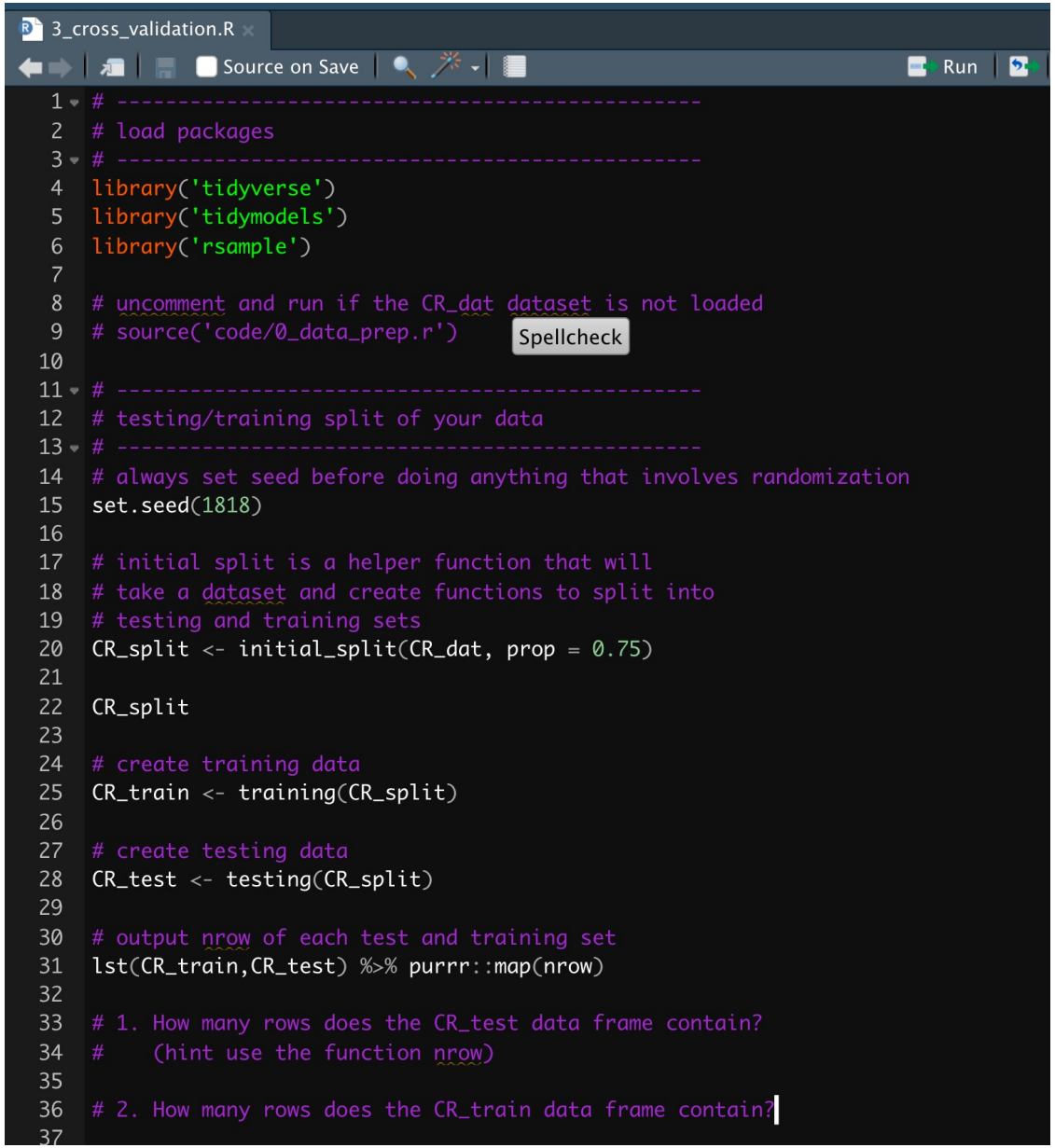
y_i	$\hat{y}_i$	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$
5	5	0	$0^2 = 0$
6	7	-1	$-1^2 = 1$
9	8	1	$1^2 = 1$
10	1	9	$9^2 = 81$

Summary – Intro to Machine Learning

- **Machine Learning** is a set of methods developed to find robust patterns across datasets
- **Public Policy 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 = bias^2 + variance

Cross-Validation Lab



- Open the file 3_cross_validation.R
- Execute the code to generate a testing and training set
- Answer the questions at the bottom