ISEA Week 5 – Machine Learning Fundamentals

Lovenoor (Lavi) Aulck

IES Institute of Education Sciences

AmplifyLearn.AI

UNIVERSITY of WASHINGTON eScience Institute
ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

UNIVERSITY OF OREGON

Outline

- Level-setting
- Model Selection
- Warm Up Exercise
- Machine Learning Overview
- Basics
- Programming Exercise

Who Am I?

IES Institute of Education Sciences

AmplifyLearn.AI

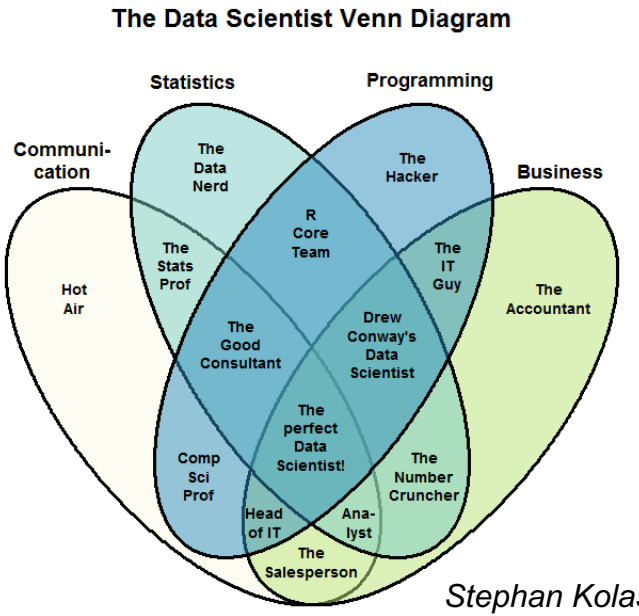
UNIVERSITY of WASHINGTON eScience Institute
ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

UNIVERSITY OF OREGON

You Will Need...

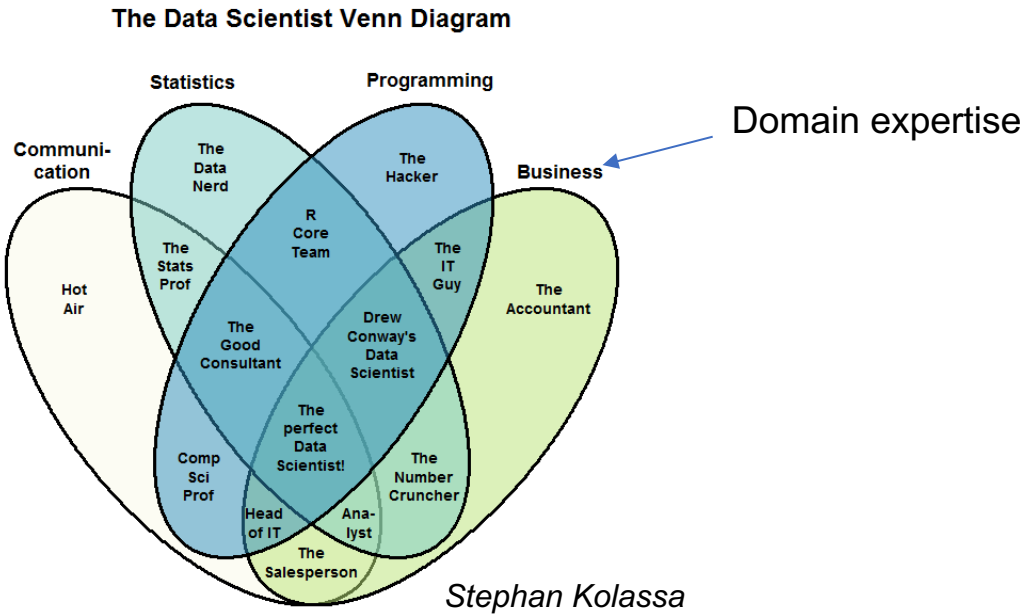
- You will need to have the following python packages installed and working for the demos:
  - pandas
  - numpy
  - seaborn (and/or matplotlib)
  - sklearn
  - statsmodels
  - random

A Data Scientist

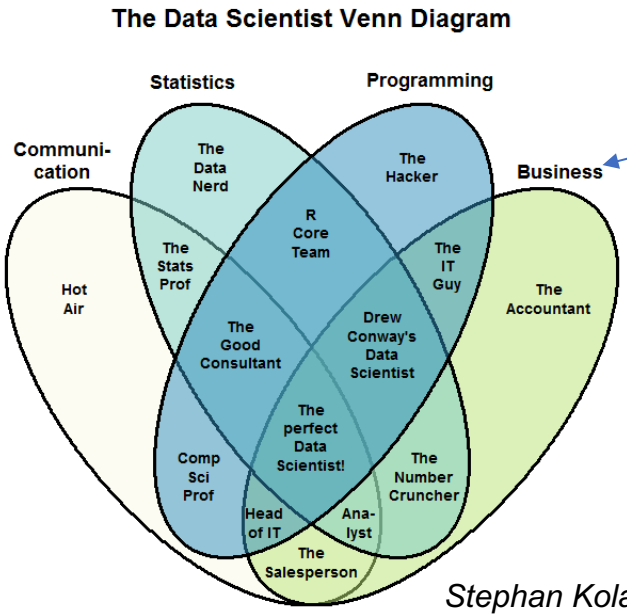


Stephan Kolassa

A Data Scientist



A Data Scientist

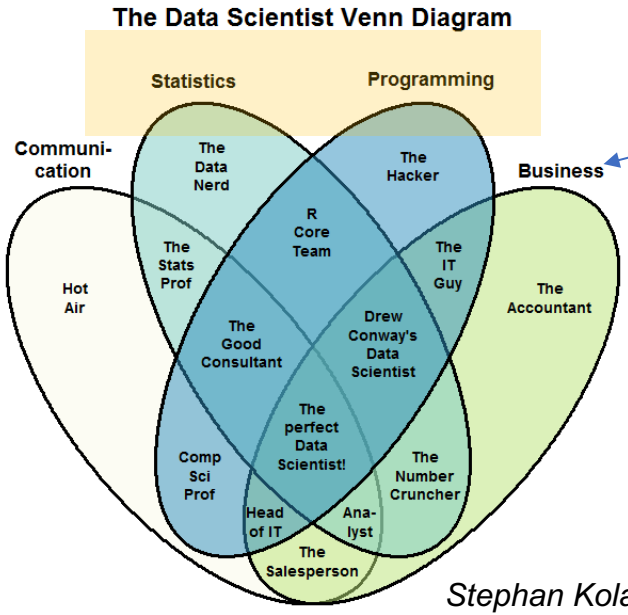


Domain expertise

Critical thinking, creativity, and experimental design?

Stephan Kolassa

A Data Scientist

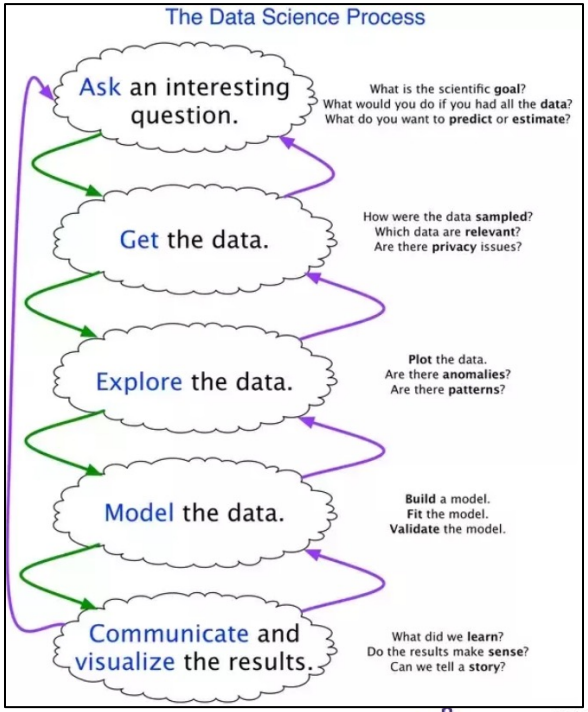


Domain expertise

Critical thinking, creativity, and experimental design?

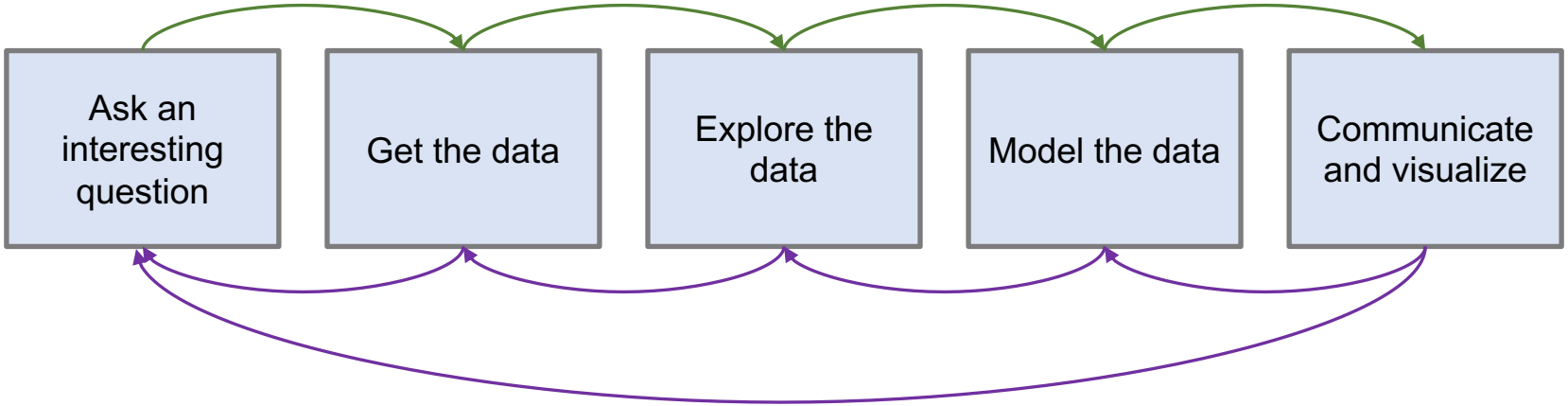
Stephan Kolassa

A Data Scientist

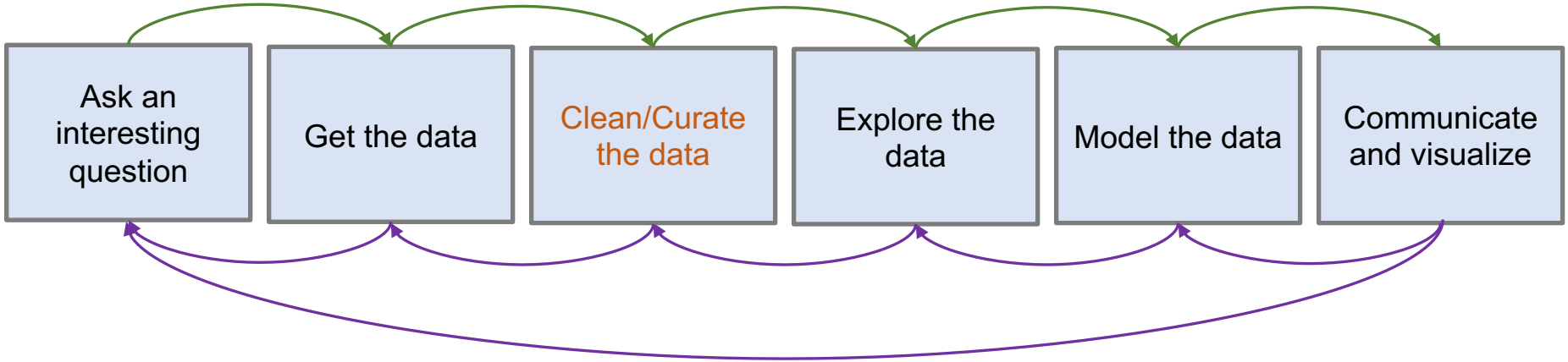


Joe Blitzstein

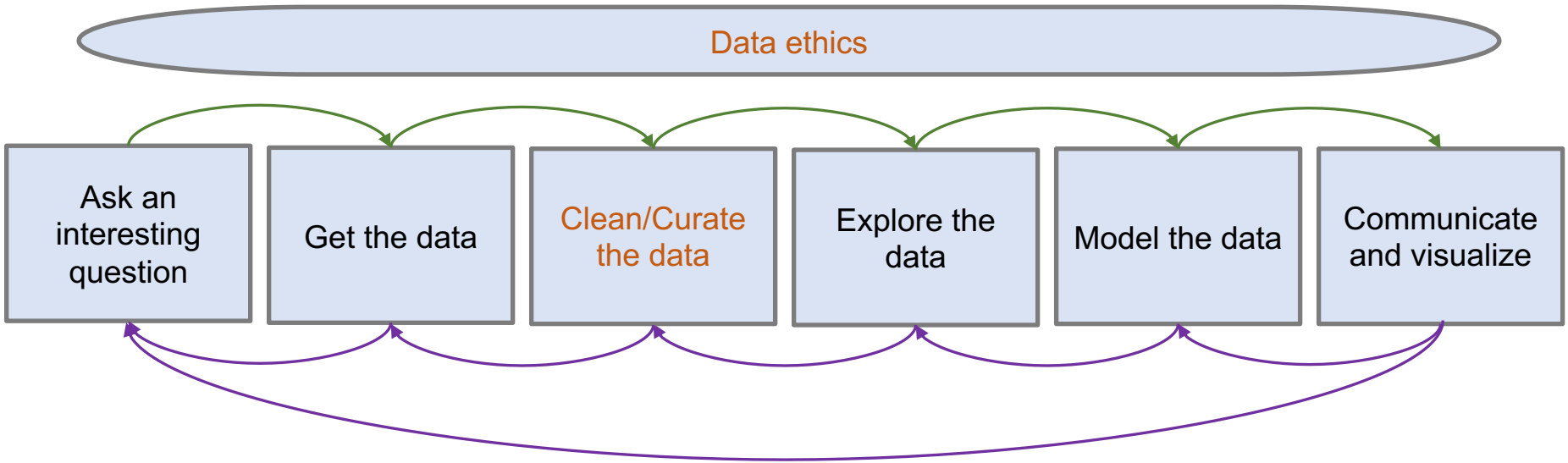
A Data Scientist



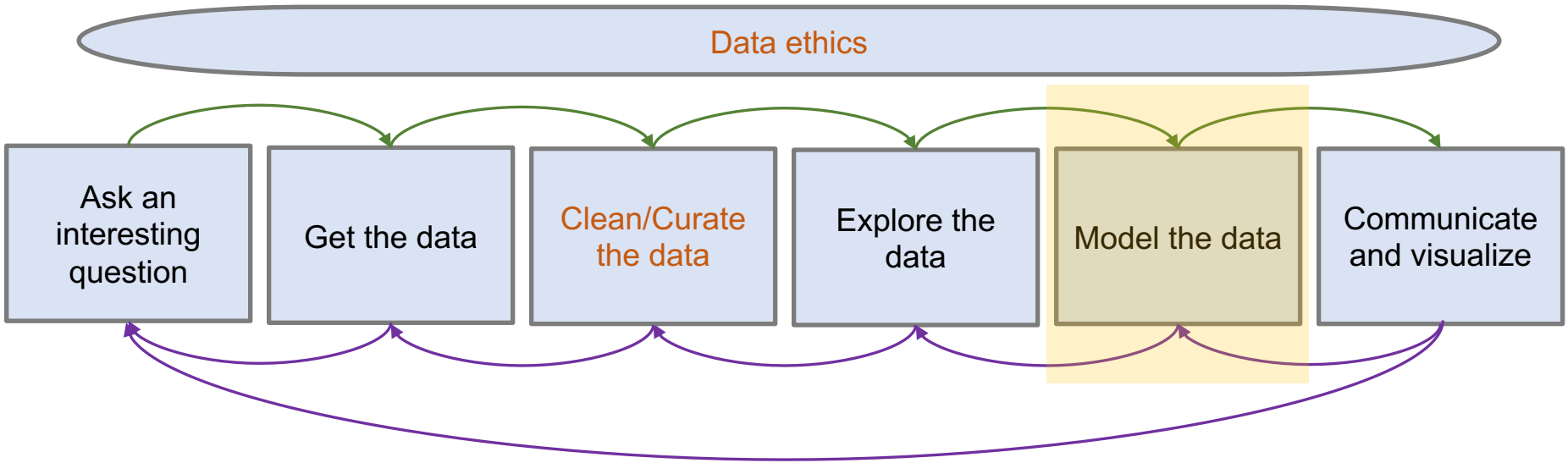
A Data Scientist



A Data Scientist



A Data Scientist



How To Build A Model (A Game)

Examine outputs from regression model

Build a regression model using explanatory variables of interest

Expand and iterate on model

Clean/curate the data

Explore the data and calculate descriptive statistics of the variables in the dataset

Examine outputs from expanded regression model

We specify a question and ensure we have the data to answer the question

How To Build A Model (A Game)

We specify a question and ensure we have the data to answer the question

Clean/curate the data

Explore the data and calculate descriptive statistics of the variables in the dataset

Build a regression model using explanatory variables of interest

Examine outputs from regression model

Expand and iterate on model

Examine outputs from expanded regression model

Multiple Paradigms of Regression

- Statistics/economics: explain a relationship
  - Ties in with the idea of inference

Multiple Paradigms of Regression

- Statistics/economics: explain a relationship
  - Ties in with the idea of inference

- Machine learning: predict the future
  - Ties in with the idea of prediction

Example: Earnings

- Inference
  - What are the determinants of income?
  - Do people with children earn more?
  - On average, how much more will a person earn for each additional year of schooling?

Example: Earnings

- Inference
  - What are the determinants of income?
  - Do people with children earn more?
  - On average, how much more will a person earn for each additional year of schooling?

- Prediction
  - What is the predicted income for person X?
  - What are the descriptors of a person with income Y?

Example: Earnings

Data on Swiss labor force outcomes:

	working	income	age	educ	kids	oldkids	foreign	welfare
1	No	48411.71	30	8	1	1	no	FALSE
2	Yes	37206.96	45	8	0	1	no	TRUE
3	no	58022.03	46	9	0	0	no	FALSE
4	no	66502.78	31	11	2	0	no	FALSE
5	no	66734.01	44	12	0	2	no	FALSE
6	yes	61589.95	42	12	0	1	no	FALSE
7	no	94344.38	51	8	0	0	no	FALSE
8	yes	35987.18	32	8	0	2	no	TRUE
9	no	41140.16	39	12	0	0	no	FALSE
10	no	35825.67	43	11	0	2	no	TRUE
11	no	42642.98	45	11	0	2	no	FALSE
12	no	35157.35	60	12	0	0	no	TRUE
13	no	75326.86	33	11	2	0	no	FALSE
14	no	148221.3	56	14	0	0	no	FALSE
15	no	98870.09	56	11	0	0	no	FALSE
16	no	80297.36	47	11	0	1	no	FALSE
17	no	52117.34	50	8	0	0	no	FALSE

Example: Earnings

- Linear regression offers a concise summary of the mean of one variable as a function of the other variable(s) through two parameters: the slope and intercept of the line

Example: Earnings

- Linear regression offers a concise summary of the mean of one variable as a function of the other variable(s) through two parameters: the slope and intercept of the line

- Model:

$$wages_i = \beta_0 + \beta_1 * education_i + \epsilon_i$$

$$wages_i = 12409 + 2518 * education_i + \epsilon_i$$

Example: Earnings

- Linear regression offers a concise summary of the mean of one variable as a function of the other variable(s) through two parameters: the slope and intercept of the line

- Model:

$$wages_i = \beta_0 + \beta_1 * education_i + \epsilon_i \qquad wages_i = 12409 + 2518 * education_i + \epsilon_i$$

- Have we measured the causal effect of an additional year of education on wages?

Example: Earnings

- Linear regression offers a concise summary of the mean of one variable as a function of the other variable(s) through two parameters: the slope and intercept of the line

- Model:

$$wages_i = \beta_0 + \beta_1 * education_i + \epsilon_i \qquad wages_i = 12409 + 2518 * education_i + \epsilon_i$$

- Have we measured the causal effect of an additional year of education on wages?
  - With another year of education, will I definitely earn $2,518 more?

The Model Is Right...?

> “All models are wrong but some are useful”

- George Box

> “... all models are limited by the validity of the assumptions on which they ride”

- Collier, Sekhon, and Stark

> “Assumptions behind models are rarely articulated, let alone defended.”

- David Freedman

Models for Predictions

- All models are, in small or large part, wrong
  - Miss features, omit dependencies, make assumptions, etc

Models for Predictions

- All models are, in small or large part, wrong
  - Miss features, omit dependencies, make assumptions, etc

- Models provide a simplification
  - Require assumptions
  - Require some understanding/intuition

Models for Predictions

- All models are, in small or large part, wrong
  - Miss features, omit dependencies, make assumptions, etc
- Models provide a simplification
  - Require assumptions
  - Require some understanding/intuition
- But… what do we need for prediction?

Models for Predictions

- But… what do we need for prediction?
  - If all we care about is predicting some target variable, maybe we can just ignore some of the messy assumptions and focus on specific metrics?

Models for Predictions

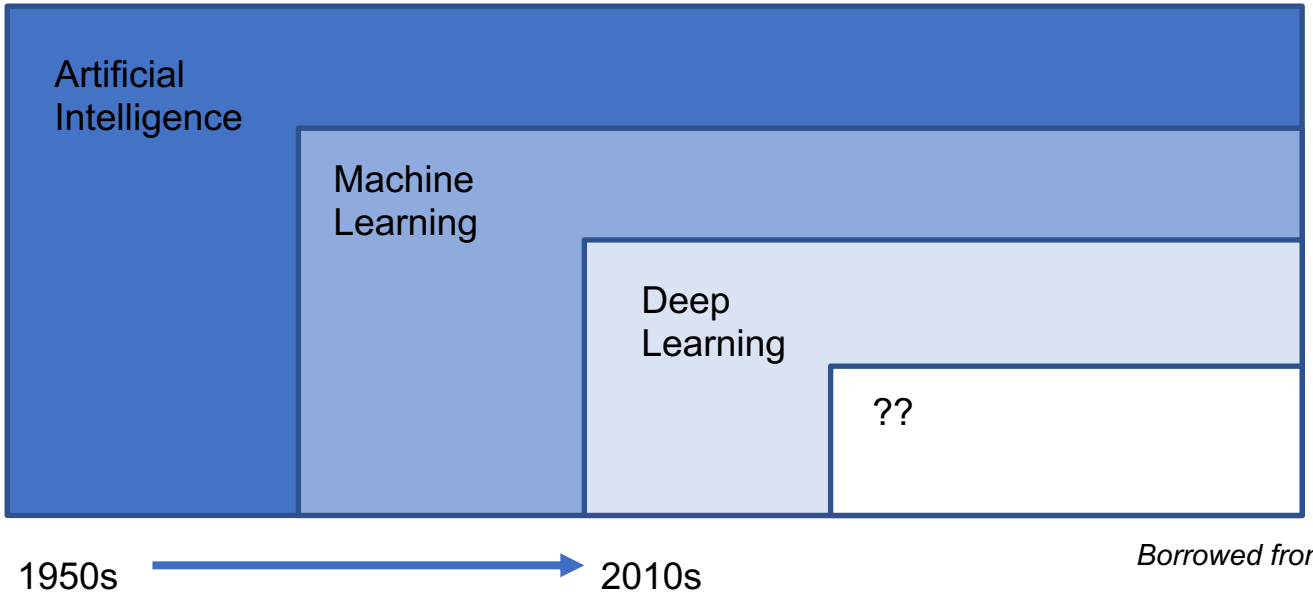
- But… what do we need for prediction?
  - If all we care about is predicting some target variable, maybe we can just ignore some of the messy assumptions and focus on specific metrics?

- A critical concern of machine learning is the ability to build models that accurately generalize while a critical concern of econometrics is the ability to build models that capture relationships

What is Machine Learning?

- Machine learning is the science of getting computers to act without being explicitly programmed
- Machine learning is a scientific discipline that explores the construction and study of algorithms that learn from data
- Machine learning is a natural outgrowth at the intersection of computer science and statistics

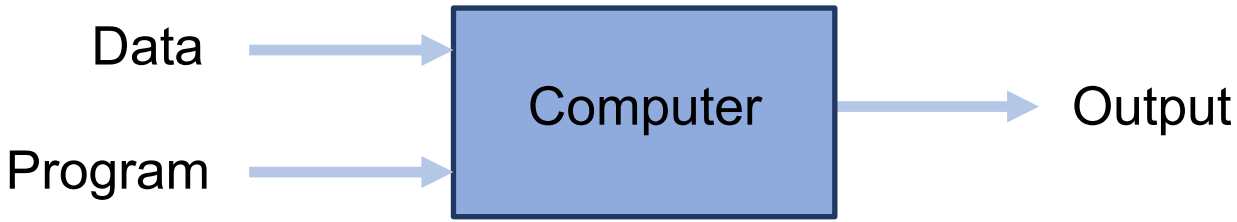
What is Machine Learning?



Borrowed from Nvidia

What is Machine Learning?

Traditional Programming

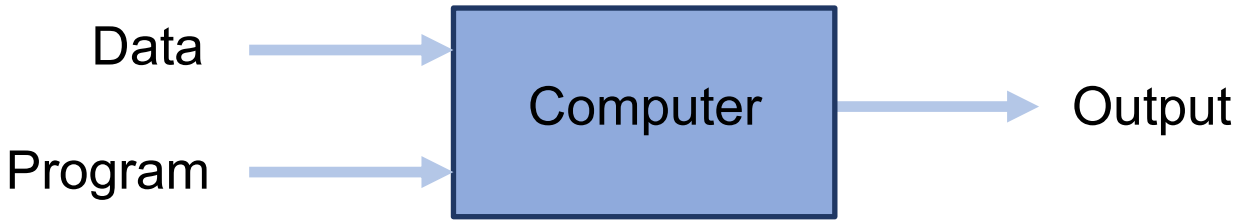


Machine Learning

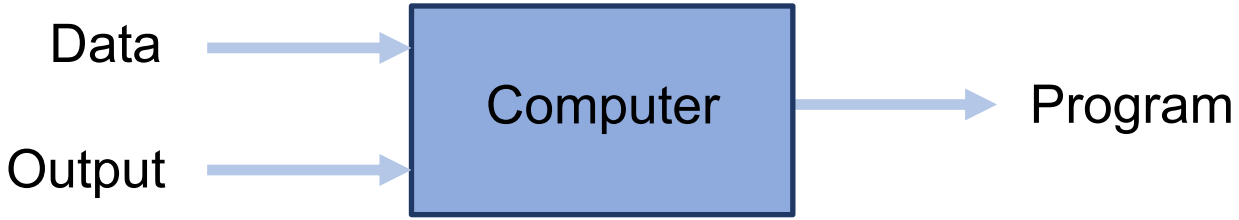
?

What is Machine Learning?

Traditional Programming



Machine Learning



What is Machine Learning?

"Learning is (like) farming, which lets nature do most of the work. Farmers combine seeds and nutrients to grow crops. Learners combine knowledge with data to grow programs."

- Pedro Domingos

“Flavors” of Machine Learning

- Supervised learning
- Unsupervised learning
- Semi-supervised learning
- Reinforcement learning

Supervised Learning

- We know the “correct” answer for values
  - We’re trying to model the input (independent) variables as they relate to the output (dependent) variables
- Examples
  - Predicting whether a student will graduate from a university
  - Predicting salary upon graduation
  - Other examples?

Unsupervised Learning

- We don't know the "correct" answer for values
  - We're trying to discover underlying structure
- Examples
  - Extracting topics/themes from textual surveys of students
  - Understanding the relationships between different departments/programs
  - Finding bottlenecks in student course offerings

“Pop” Quiz

Would you address each of the below with a supervised or unsupervised learning algorithm?

- Given email labelled spam or not, build a spam detector
- Given news articles on the web, group them into sets based on topic
- Given a database of customer data, find market segments/customer groups
- Given patients with a disease, determine if new patients have the disease
- Given phone records of individuals, determine which are the wealthiest
- Given phone records of individuals and survey data about their income, predict the incomes of new subscribers

Semi-supervised Learning

- We know the “correct” answer for some values
  - The input data contains both labelled and unlabeled instances
- Examples?

Reinforcement Learning

- We develop “agents” that seek to maximize “reward”
  - We’re trying to build models that will seek to maximize some gain
- Examples
  - Creating personalized curricula for students
  - Creating educational content for students

ML Basics

- Thousands of ML algorithms
  - No one knows them all
  - Often tweaks or small improvements to existing approaches

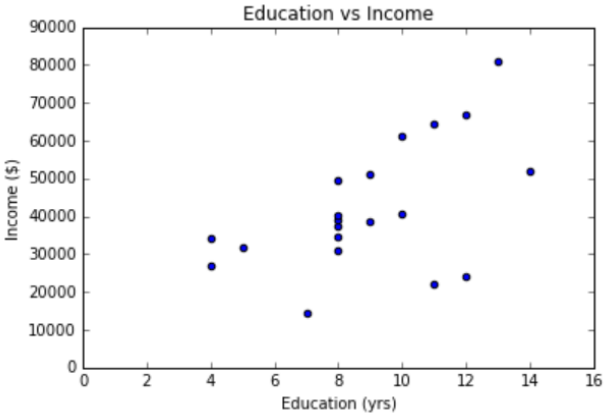
ML Basics

- Thousands of ML algorithms
  - No one knows them all
  - Often tweaks or small improvements to existing approaches
- Every ML algorithm has three components
  - The representation (the model)
  - The evaluation (the cost/objective function)
  - The optimization (the search)

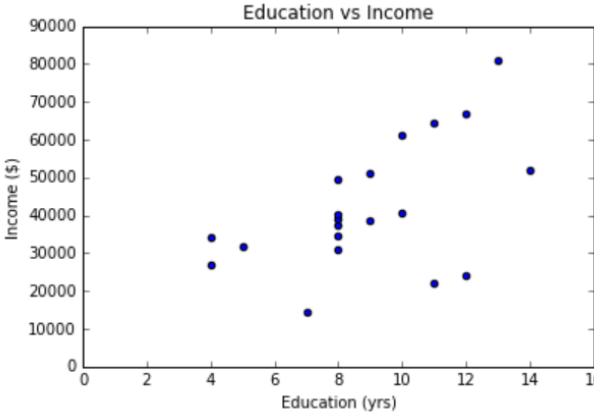
Designing for Prediction: Key Ideas

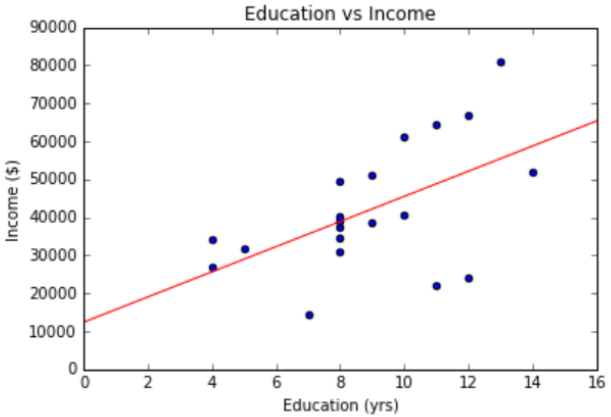
- Generalization and overfitting
- Training, validation, and test data
- Evaluation metrics
- Baselines
- Error analysis

Modeling Education and Income

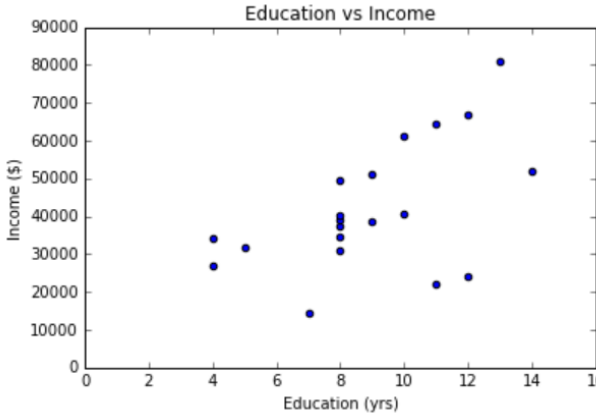


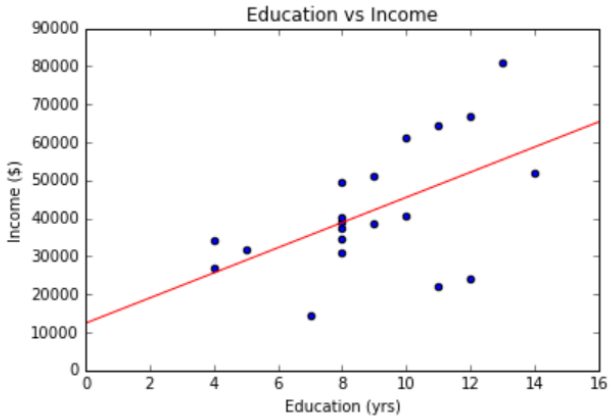
Modeling Education and Income

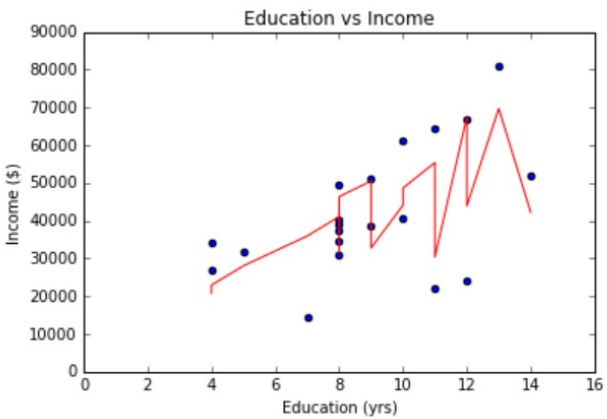




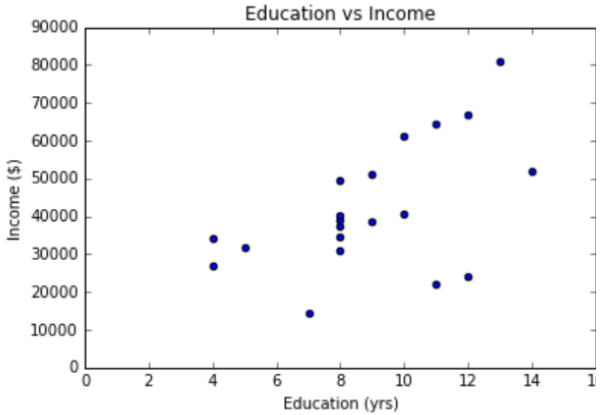
Modeling Education and Income



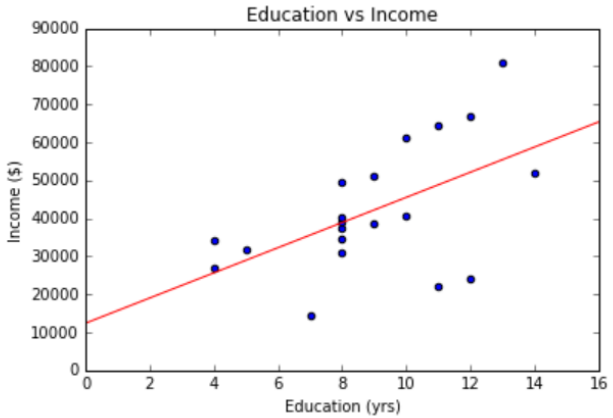




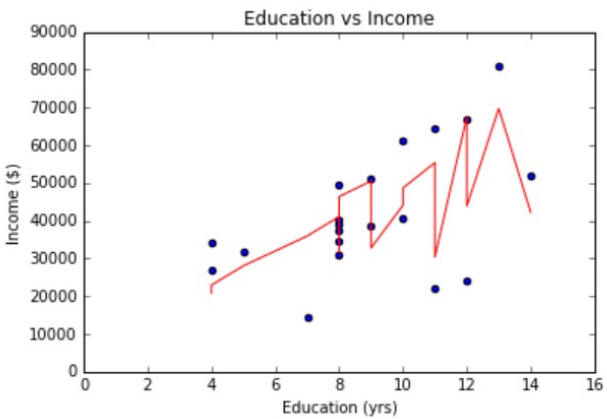
Modeling Education and Income



Plotting

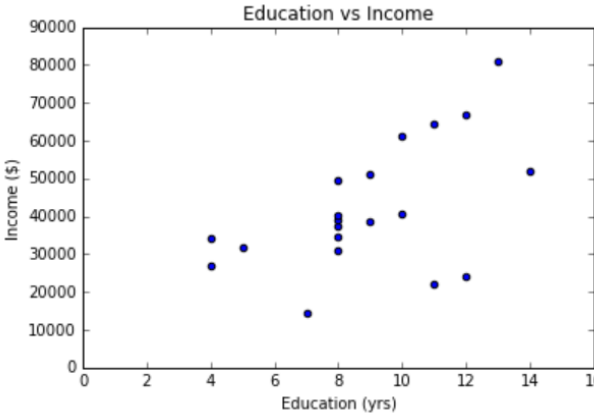


One Variable

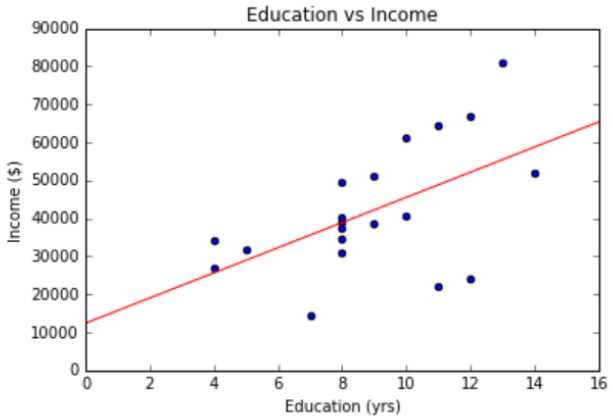


“Kitchen Sink”

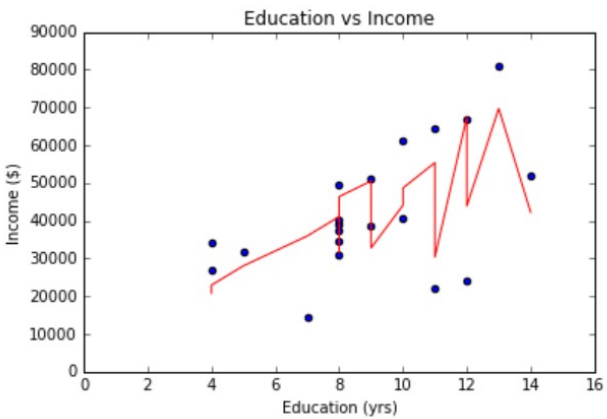
Modeling Education and Income



Plotting



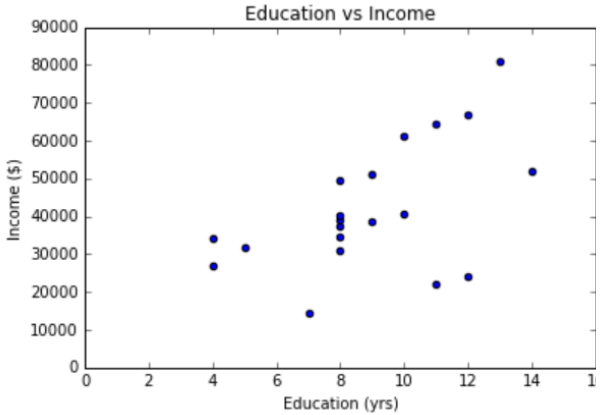
One Variable



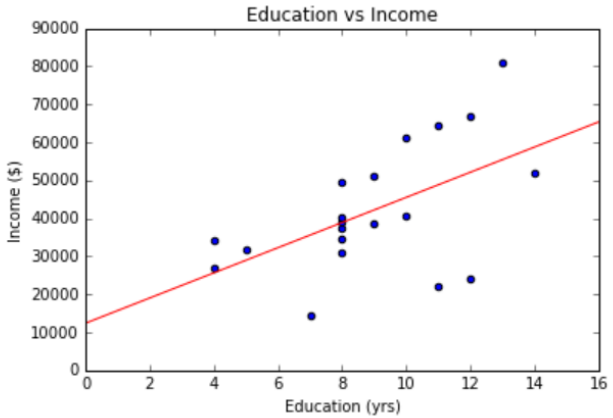
"Kitchen Sink"

Which of these would you trust for predictions when education is 8 years?

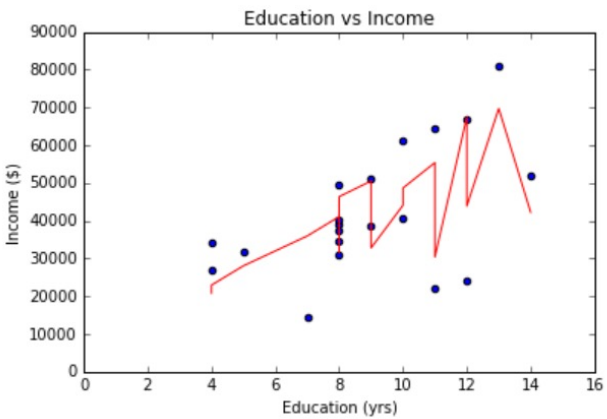
Modeling Education and Income



Plotting



One Variable



"Kitchen Sink"

Which of these would you trust for predictions when education is 8 years? What about 16 years? What about 0 years?

Notebook #1

Determining Model Fit

- Adding more features will generally improve the “fit” of your model

Determining Model Fit

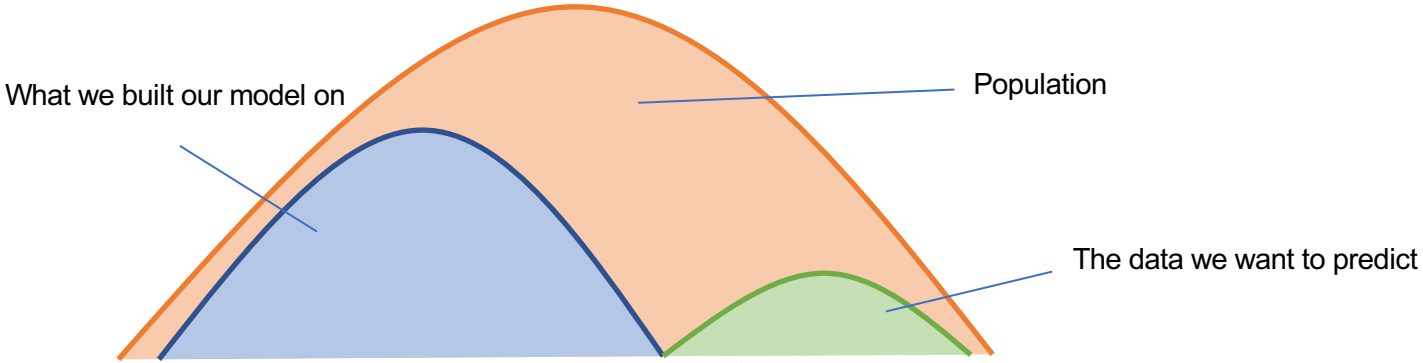
- Adding more features will generally improve the “fit” of your model
- As the number of features we have approaches or exceeds the number of observations, the model can fit really, really well
  - Of course, M > N can also lead to non-unique solutions!

Determining Model Fit

- Adding more features will generally improve the “fit” of your model
- As the number of features we have approaches or exceeds the number of observations, the model can fit really, really well
  - Of course, M > N can also lead to non-unique solutions!
- Should we keep adding features? Is this what we want?

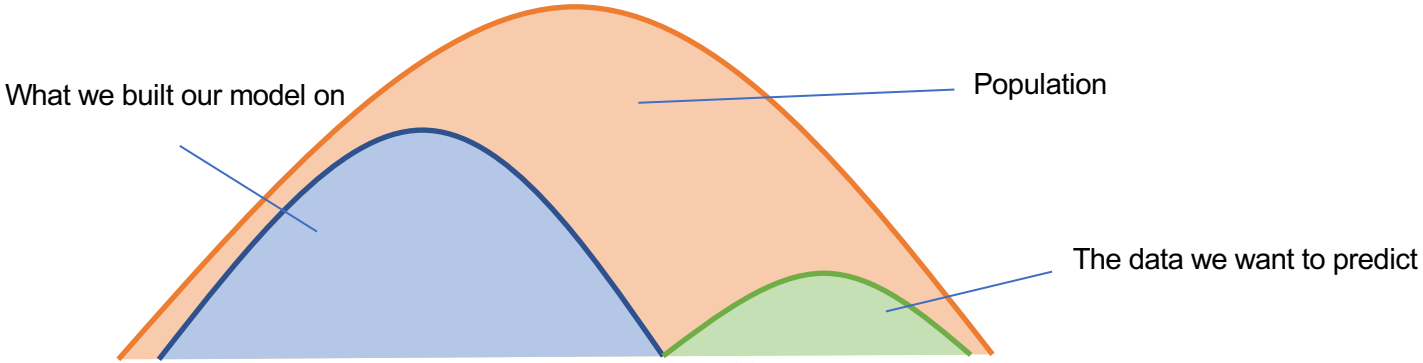
The ML Dilemma

> We want to find a balance between modeling the data we have while also being able to generalize to unseen data



The ML Dilemma

> We want to find a balance between modeling the data we have while also being able to generalize to unseen data

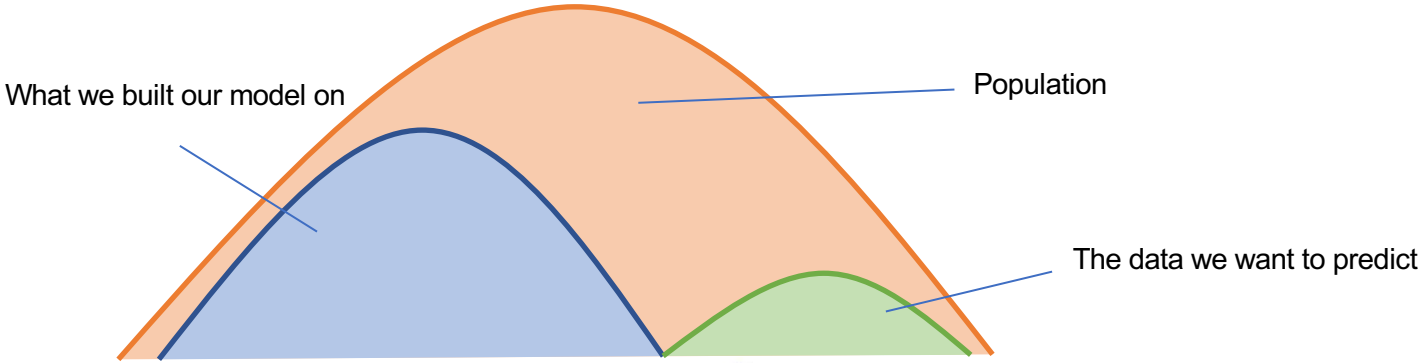


The ML Dilemma

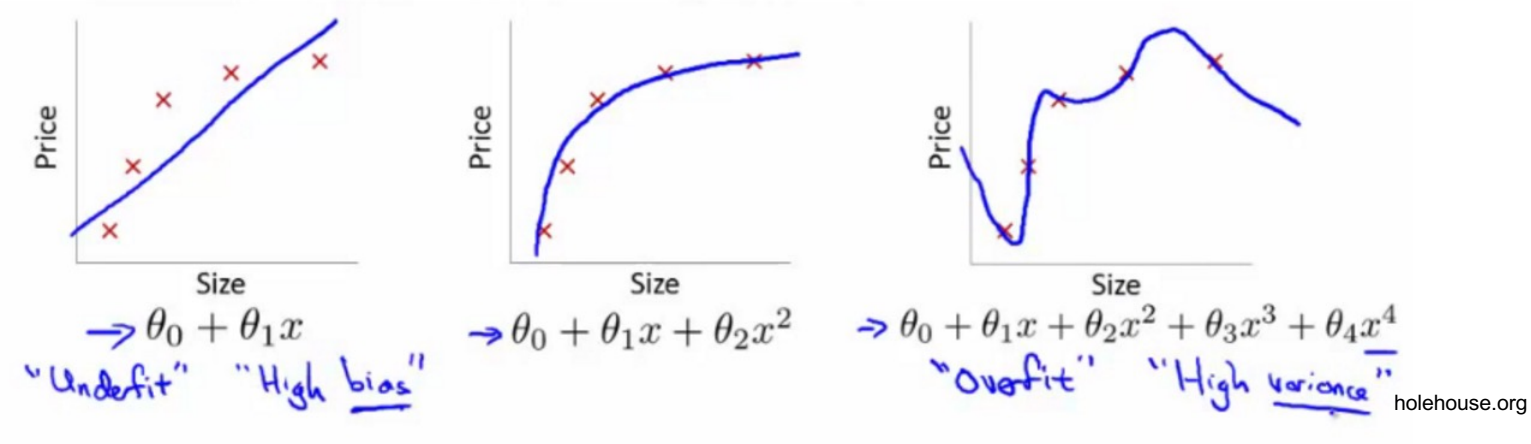
> We want to find a balance between modeling the data we have while also being able to generalize to unseen data

variance

bias



Bias-Variance Tradeoff



Bias-Variance Tradeoff: the problem of simultaneously minimizing two sources of error that prevent generalization

Overfitting and Underfitting

- More formally: overfitting is when we have too high of variance in our model with respect to bias

- More formally: underfitting is when we have too high of bias in our model with respect to variance

Overfitting and Underfitting

- More formally: overfitting is when we have too high of variance in our model with respect to bias
- More formally: underfitting is when we have too high of bias in our model with respect to variance
- How do we know?

Overfitting and Underfitting

- More formally: overfitting is when we have too high of variance in our model with respect to bias
- More formally: underfitting is when we have too high of bias in our model with respect to variance
- How do we know? Validation!

Training, Validation, and Test Data

- **Training set**: a set of examples used for learning, to which we fit parameters of a model

Training, Validation, and Test Data

- **Training set**: a set of examples used for learning, to which we fit parameters of a model
- **Validation set**: a set of examples used to tune model hyperparameters of a model (often a subset of the training set)

Training, Validation, and Test Data

- **Training set**: a set of examples used for learning, to which we fit parameters of a model
- **Validation set**: a set of examples used to tune model hyperparameters of a model (often a subset of the training set)
- **Test set**: a set of examples used only to assess the performance of a fully-specified model. DO NOT look at the test data to guide model design!!

Training, Validation, and Test Data

- **Training set**: a set of examples used for learning, to which we fit parameters of a model
- **Validation set**: a set of examples used to tune model hyperparameters of a model (often a subset of the training set)
- **Test set**: a set of examples used only to assess the performance of a fully-specified model. DO NOT look at the test data to guide model design!!

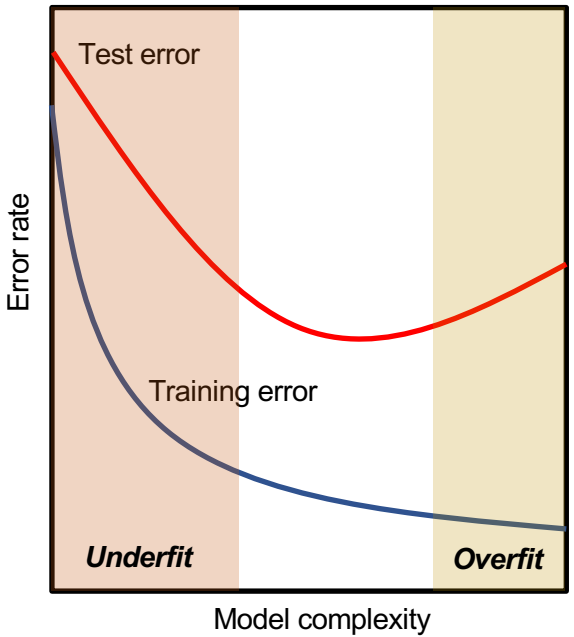
Test Data

Until you're ready to report final results...

Do not peek at your test data!! Do not peek at your test data!!
Do not peek at your test data!! Do not peek at your test data!!
Do not peek at your test data!! Do not peek at your test data!!
Do not peek at your test data!! Do not peek at your test data!!
Do not peek at your test data!! Do not peek at your test data!!
Do not peek at your test data!! Do not peek at your test data!!
Do not peek at your test data!! Do not peek at your test data!!
Do not peek at your test data!! Do not peek at your test data!!
Do not peek at your test data!! Do not peek at your test data!!
Do not peek at your test data!! Do not peek at your test data!!

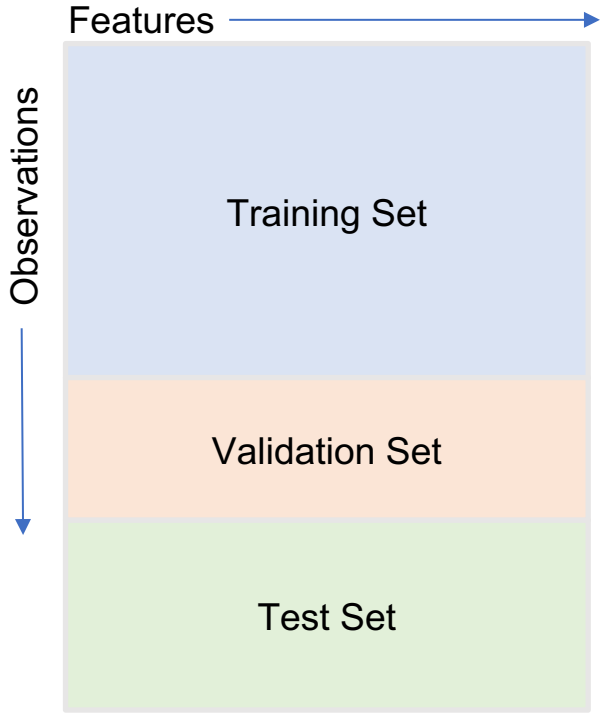
How Prediction Experiments Work

- ML experiments typically separate data into a training/test set
- Model is fit on training set
  - Validation set is often used to fine tune
- Performance is measured on the test set
  - This gives something of a "real-world" approximation of how well the model performs



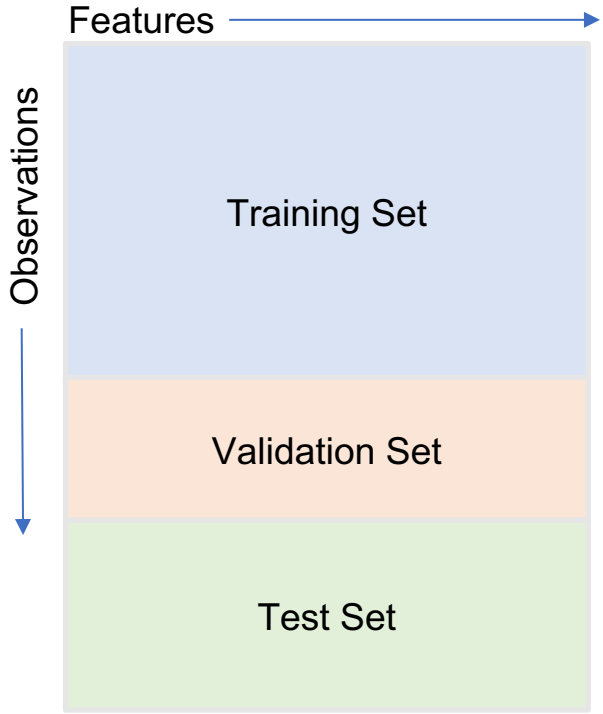
A Typical (Supervised) ML Experiment

- Data with labelled instances
  - Split into training, validation, and test sets



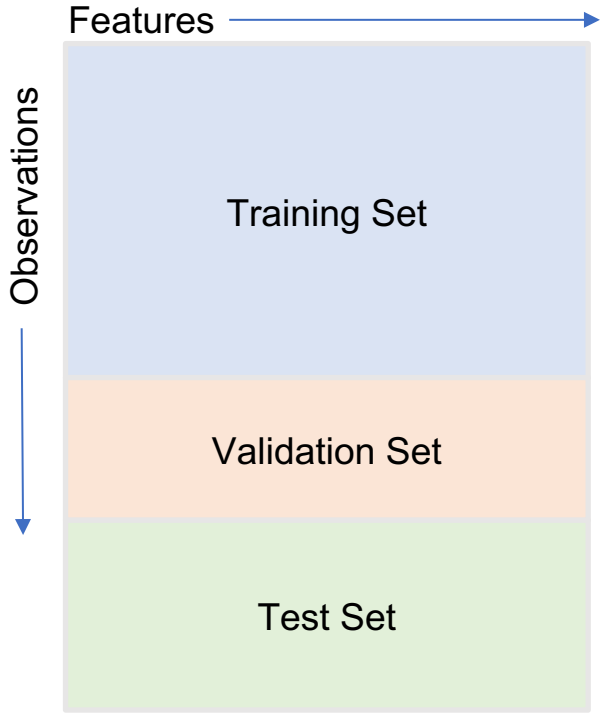
A Typical (Supervised) ML Experiment

- Data with labelled instances
  - Split into training, validation, and test sets
- Training your model
  - Estimate parameters on training set
  - Tune hyperparameters on the validation set
  - Report results on the test set



A Typical (Supervised) ML Experiment

- Data with labelled instances
  - Split into training, validation, and test sets
- Training your model
  - Estimate parameters on training set
  - Tune hyperparameters on the validation set
  - Report results on the test set
- Evaluation
  - Many metrics and context-dependent
  - Ideally, what we use to train should be our final evaluation criteria



What is a Hyperparameter?

- Parameter:
  - A value for the model that is derived via training
  - Inherent to the data
  - Example: beta coefficients in regression

What is a Hyperparameter?

- Parameter:
  - A value for the model that is derived via training
  - Inherent to the data
  - Example: beta coefficients in regression
- Hyperparameter:
  - A value that is set before the learning process
  - External to the data
  - Often tuned with a particular end result in mind
  - Example: regularization strength in regression

Bootstrapping

> Given unlimited data, it's easy to get new test data (assuming IID). What if you have limited data?

Bootstrapping

- Given unlimited data, it's easy to get new test data (assuming IID). What if you have limited data?

- Your "random" sample of training data may still not be representative
  - Remember: GIGO! (Garbage in, garbage out)

Bootstrapping

- Given unlimited data, it's easy to get new test data (assuming IID). What if you have limited data?

- Your "random" sample of training data may still not be representative
  - Remember: GIGO! (Garbage in, garbage out)

- Bootstrapping can help us recycle and maximize data usage
  - It also allows us an easy avenue to tune hyperparameters

K-Fold Cross Validation

- A form of bootstrapping
- Uses K “folds”
  - K is typically 3, 5, or 10

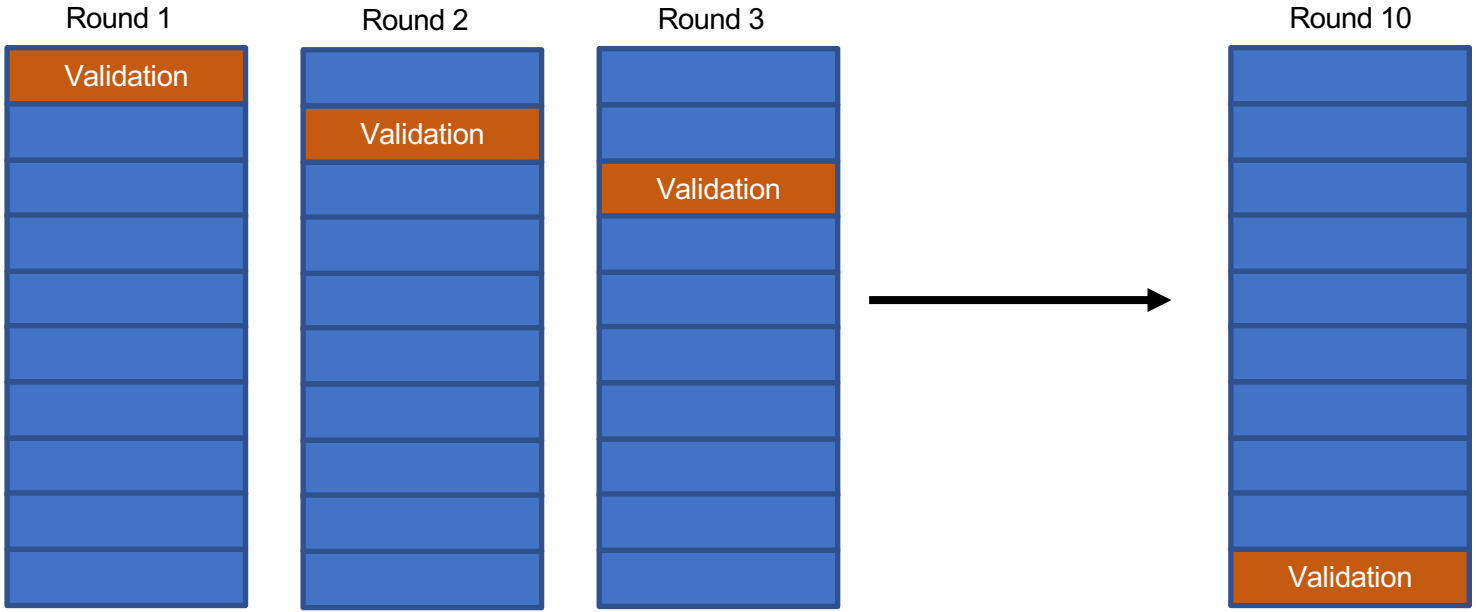
K-Fold Cross Validation

- A form of bootstrapping
- Uses K “folds”
  - K is typically 3, 5, or 10
- The process:
  - Randomly partition your (training) data into K subsamples of equal size
  - Use each of the K folds as your validation set once
  - Average your performance across the K test runs

K-Fold Cross Validation

- A form of bootstrapping
- Uses K “folds”
  - K is typically 3, 5, or 10
- The process:
  - Randomly partition your (training) data into K subsamples of equal size
  - Use each of the K folds as your validation set once
  - Average your performance across the K test runs
- Involves fitting and re-fitting the model K times!
  - Can be computationally cumbersome

K-Fold Cross Validation

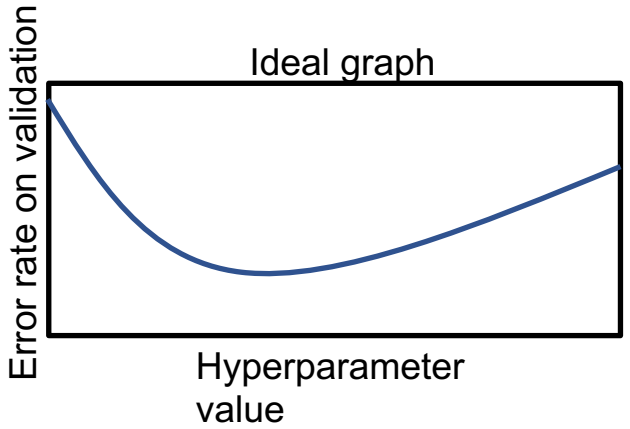


K-Fold CV to Sweep

> KFCV can be used to identify which hyperparameter settings work best

K-Fold CV to Sweep

- KFCV can be used to identify which hyperparameter settings work best
- Process:
  - Start at a value of interest for hyperparameter(s) of interest
  - Use KFCV to evaluate
  - Adjust hyperparameters
  - Repeat
- Use optimized hyperparameters to train the final model



Evaluation Metrics

- Figuring out which to use can be critical for evaluation
  - Can also weigh multiple metrics at once

Evaluation Metrics

- Figuring out which to use can be critical for evaluation
  - Can also weigh multiple metrics at once
- Classification:
  - Accuracy
  - Precision/recall
  - AUROC
- Regression:
  - MSE/RMSE

Evaluation Metrics

- Figuring out which to use can be critical for evaluation
  - Can also weigh multiple metrics at once
- Classification:
  - Accuracy
  - Precision/recall
  - AUROC
- Regression:
  - MSE/RMSE
- Log loss for either
- Many, many more

Baselines

- We also often need to quantify progress (accuracy, correctness) relative to something meaningful. The following are often used:
  - vs random guessing
  - vs most likely label
  - vs state of the art
  - vs something else simple/intuitive

Baselines

- We also often need to quantify progress (accuracy, correctness) relative to something meaningful. The following are often used:
  - vs random guessing
  - vs most likely label
  - vs state of the art
  - vs something else simple/intuitive

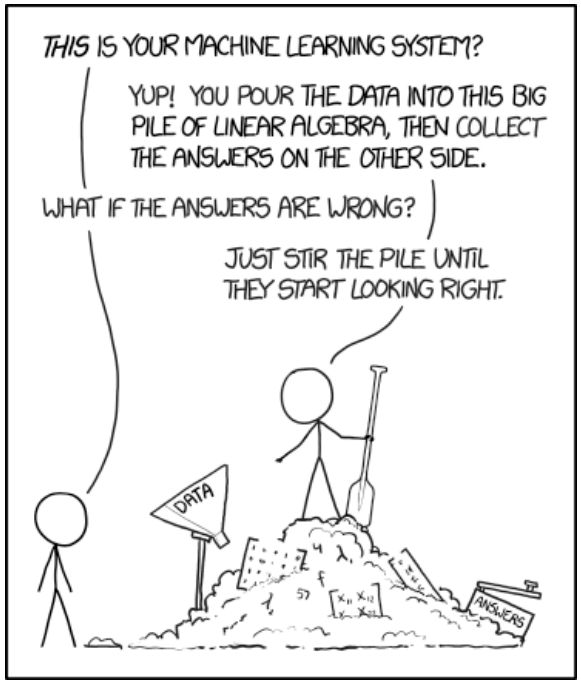
- Example: if our dataset has 90% of students retained, is an accuracy of 90% when predicting retention really all that great?

Error Analysis

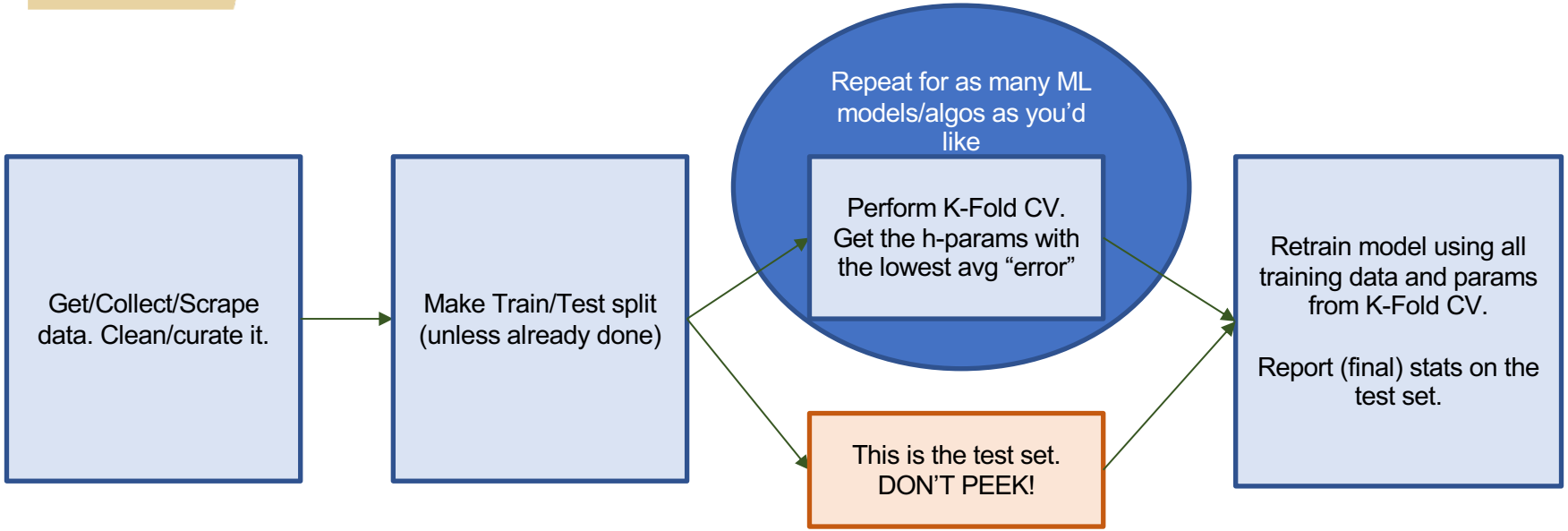
- Error analysis tries to explain the gap between current and perfect performance
  - Why is the model incorrect?
  - Where is it incorrect?
  - What types of observations does it need more leverage for?

- Things rarely work out of the box
  - Be ready to tune/tweak and re-tweak

Error Analysis



The Workflow*



*not always the workflow

Notebook #2

Ethics

> Beware of black boxes

Ethics

- Beware of black boxes
- Pay attention to “explainable” ML
  - Just because something predicts well, doesn’t mean we can’t understand it

Ethics

- Beware of black boxes
- Pay attention to “explainable” ML
  - Just because something predicts well, doesn’t mean we can’t understand it
- Pay attention to biases in your data!
  - Your models will always reflect the biases they inherit
  - This can also relate to blindspots in your data

Homework

Your homework for next week:

- On the back end of the second notebook
- Do the readings/watch the videos

Wrapping Up

- Level-setting
- Model Selection
- Warm Up Exercise
- Machine Learning Overview
- Basics
- Programming Exercise