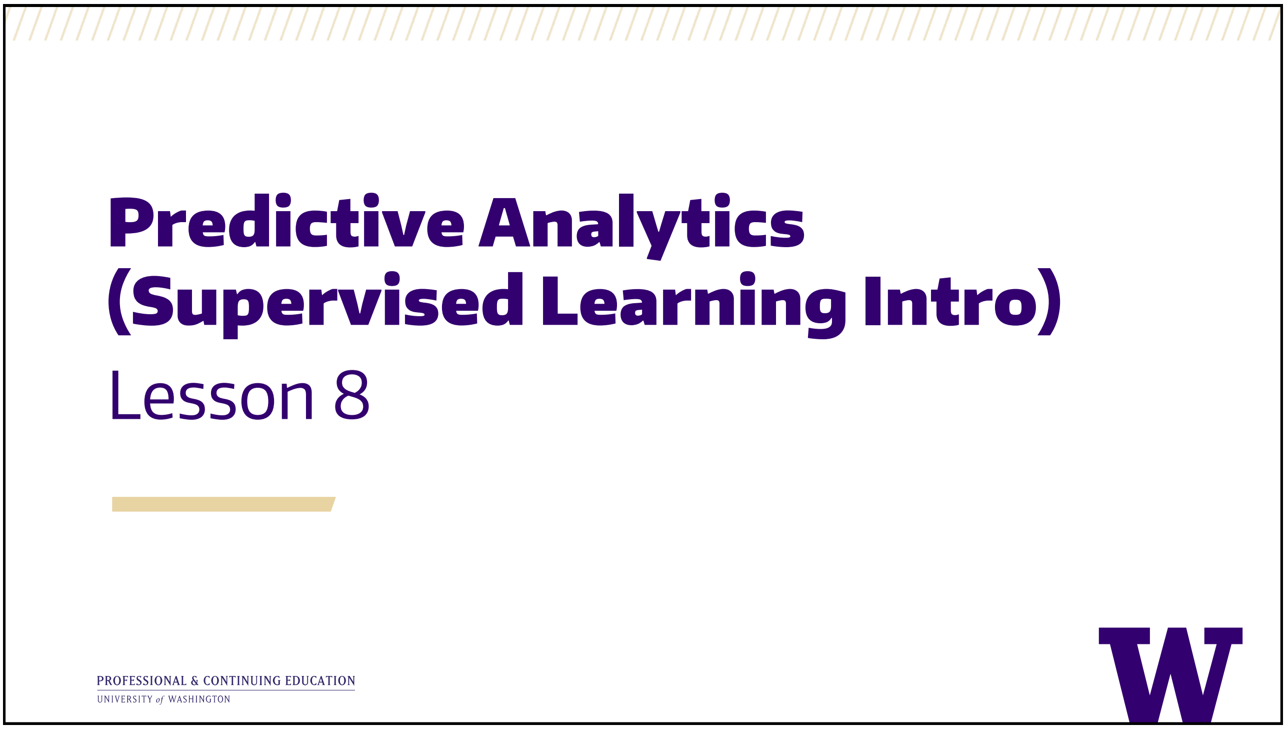
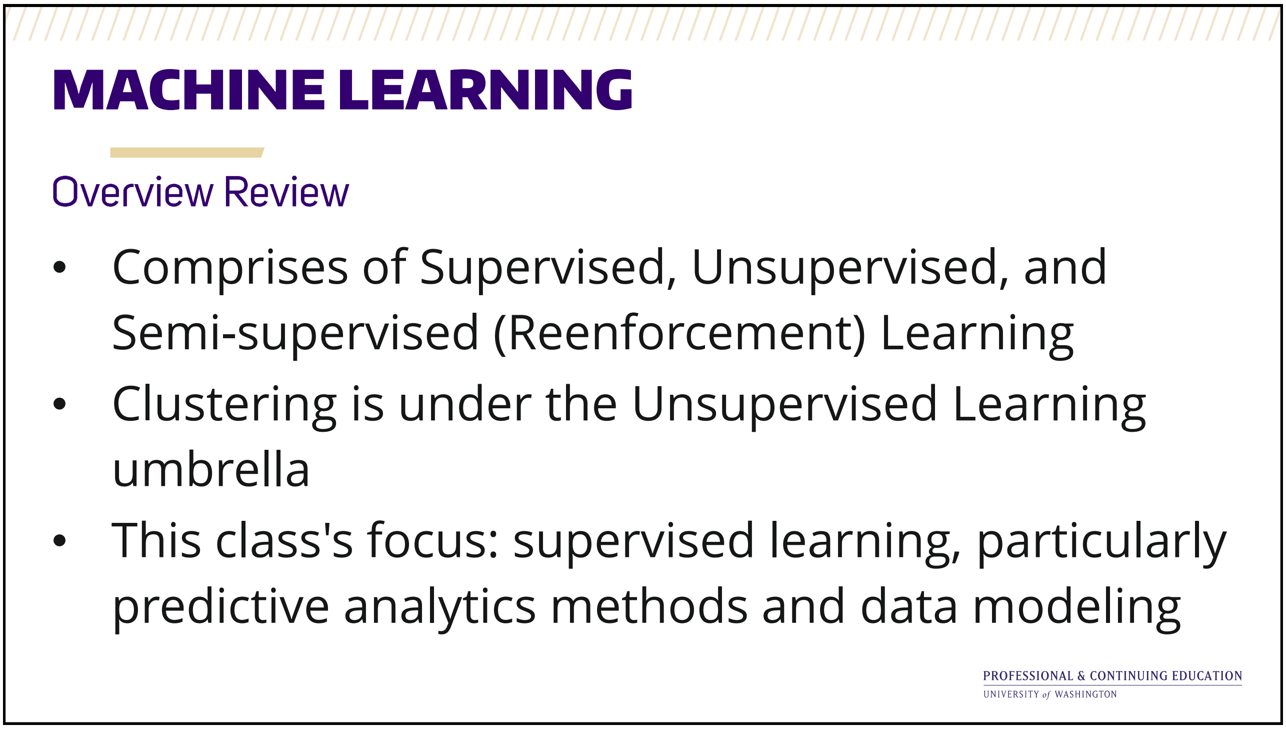


Predictive Analytics (Supervised Learning Intro)

Lesson 8

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

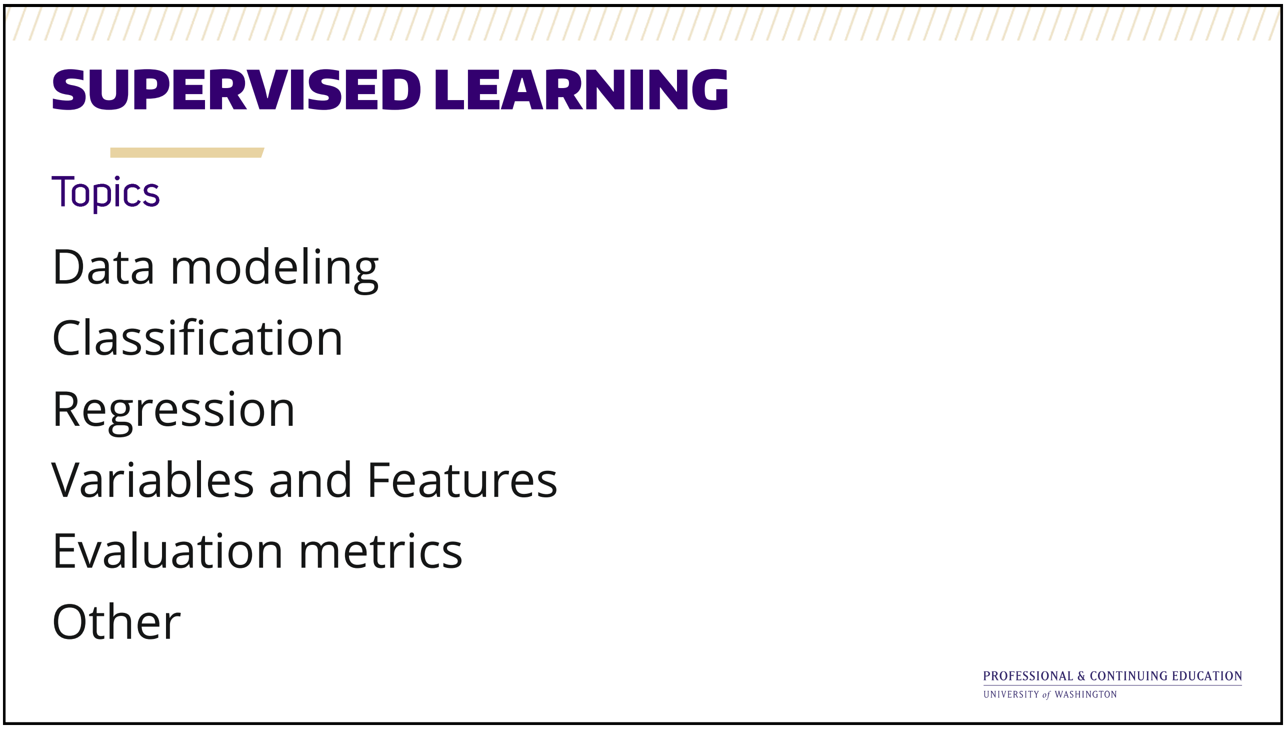


MACHINE LEARNING

Overview Review

- Comprises of Supervised, Unsupervised, and Semi-supervised (Reenforcement) Learning
- Clustering is under the Unsupervised Learning umbrella
- This class's focus: supervised learning, particularly predictive analytics methods and data modeling

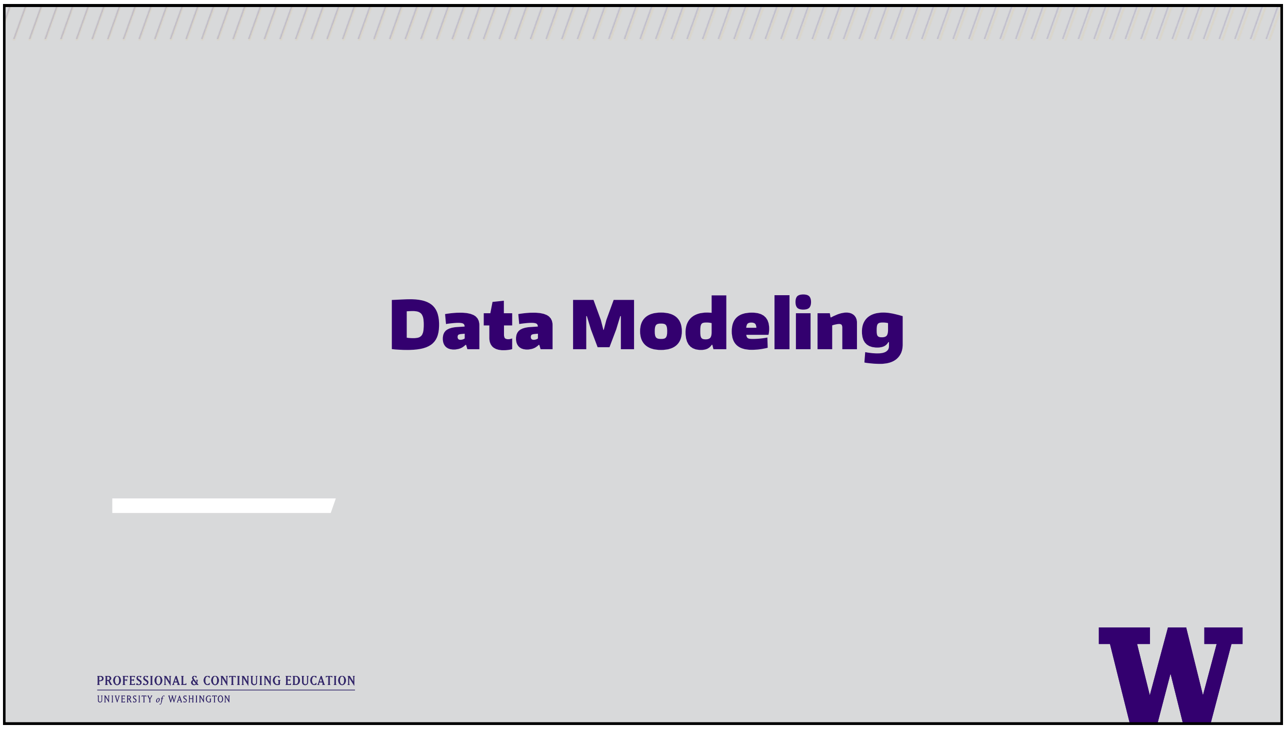
PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON



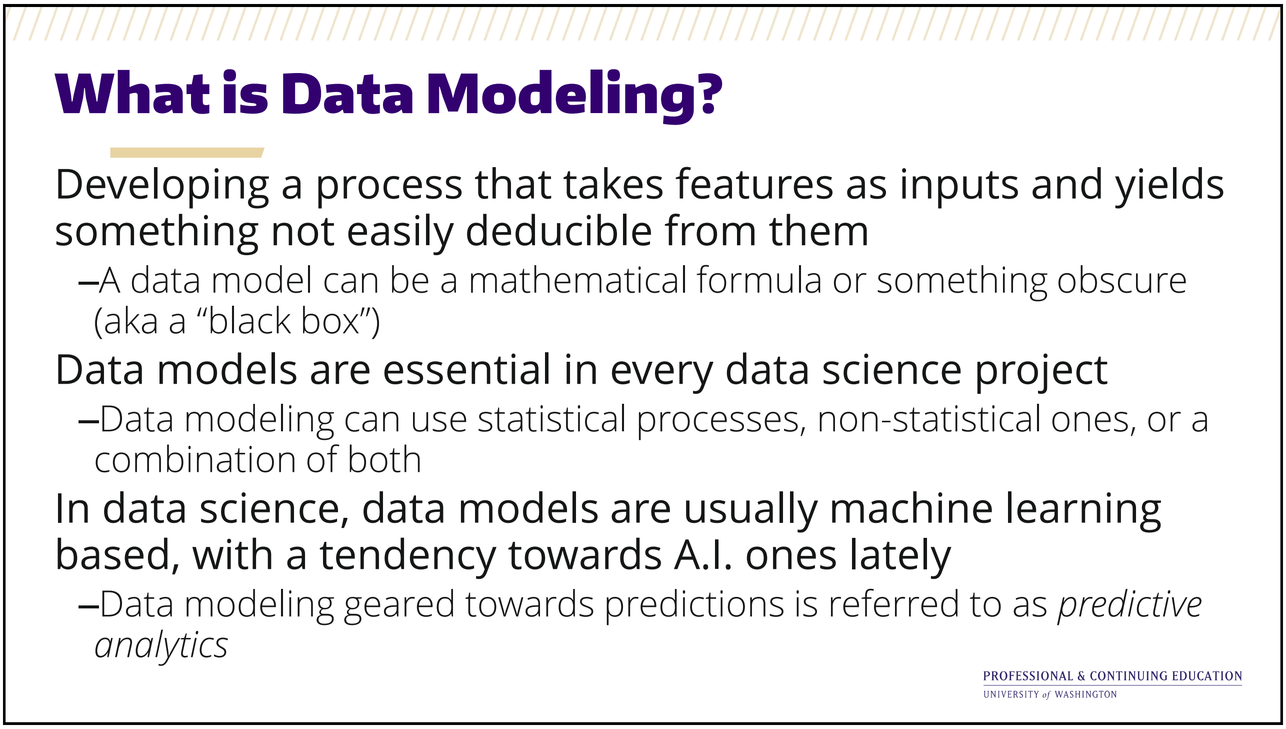
SUPERVISED LEARNING

Topics

- Data modeling
- Classification
- Regression
- Variables and Features
- Evaluation metrics
- Other



Data Modeling



What is Data Modeling?

Developing a process that takes features as inputs and yields something not easily deducible from them

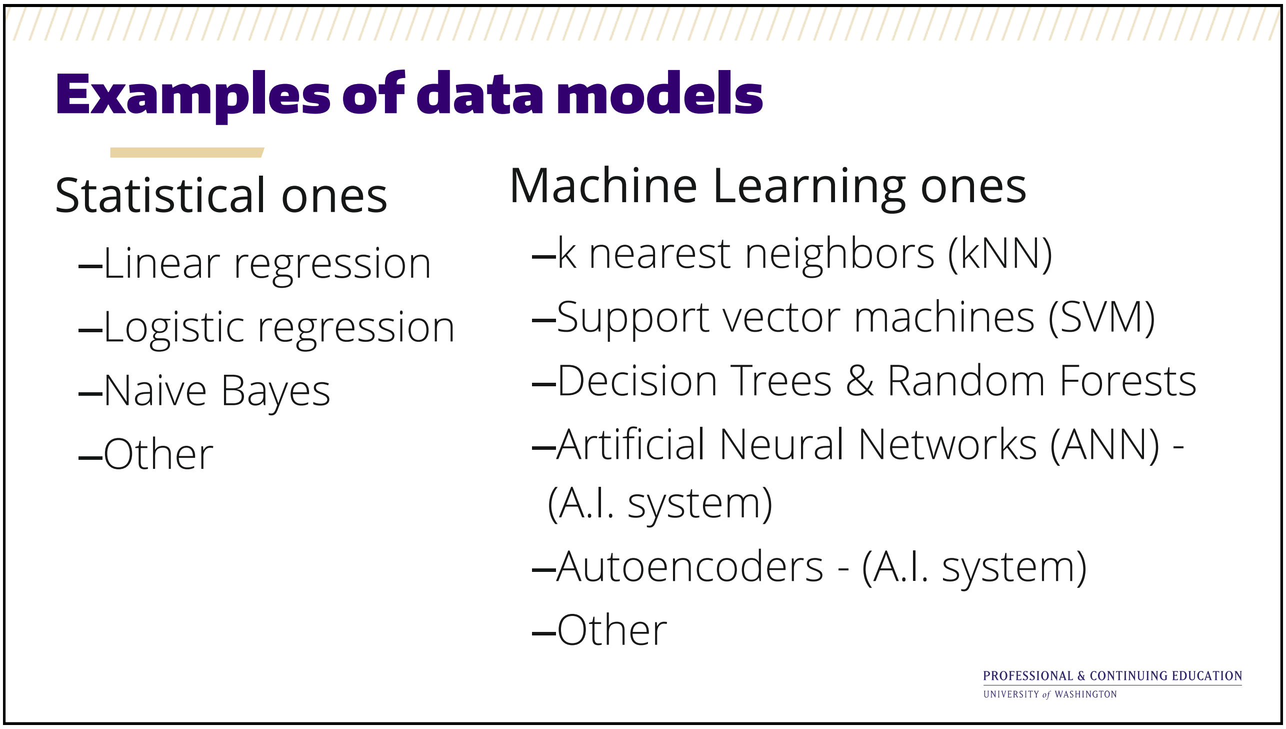
- A data model can be a mathematical formula or something obscure (aka a “black box”)

Data models are essential in every data science project

- Data modeling can use statistical processes, non-statistical ones, or a combination of both

In data science, data models are usually machine learning based, with a tendency towards A.I. ones lately

- Data modeling geared towards predictions is referred to as *predictive analytics*



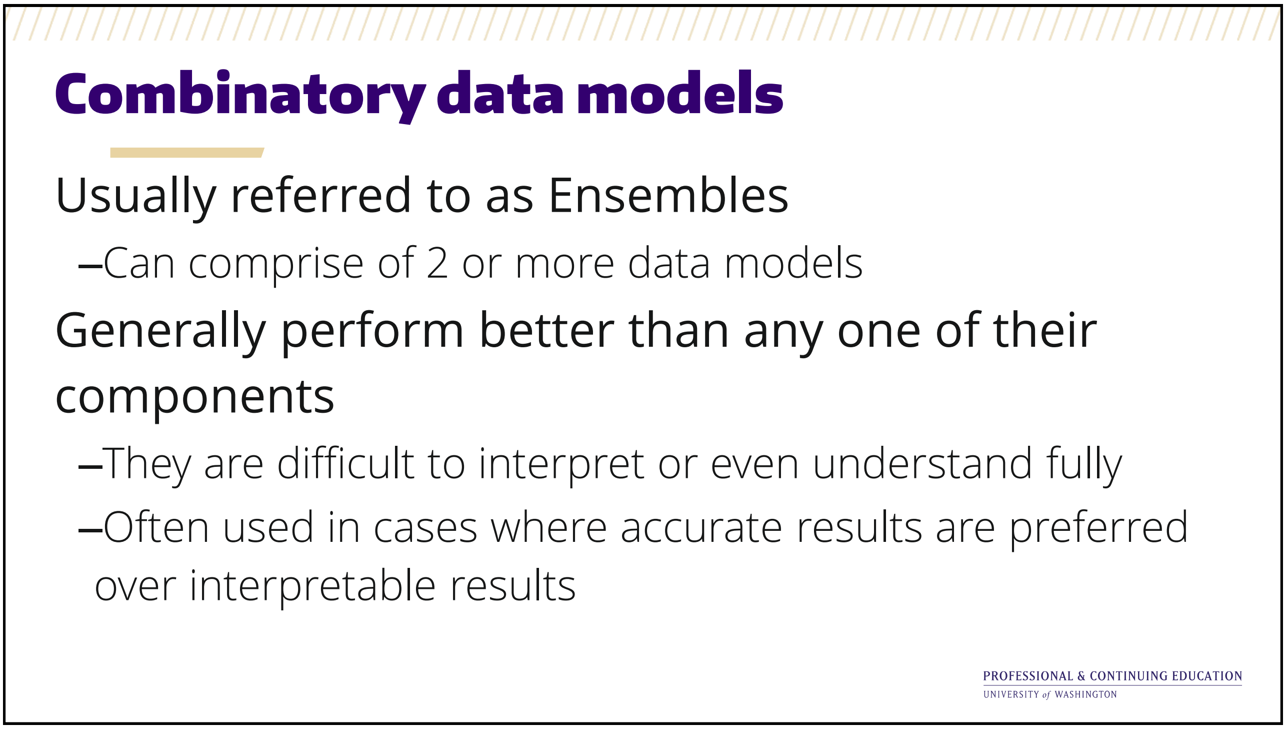
Examples of data models

Statistical ones

- Linear regression
- Logistic regression
- Naive Bayes
- Other

Machine Learning ones

- k nearest neighbors (kNN)
- Support vector machines (SVM)
- Decision Trees & Random Forests
- Artificial Neural Networks (ANN) - (A.I. system)
- Autoencoders - (A.I. system)
- Other



Combinatory data models

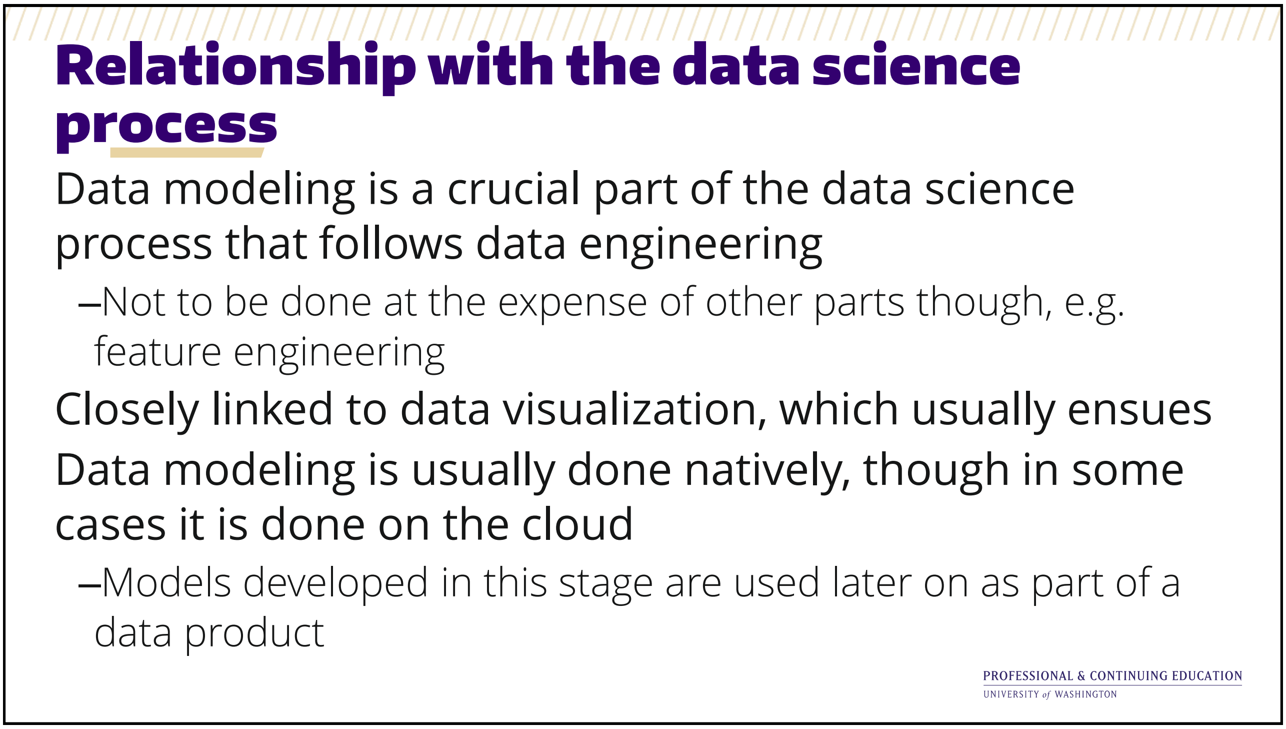
Usually referred to as Ensembles

- Can comprise of 2 or more data models

Generally perform better than any one of their components

- They are difficult to interpret or even understand fully
- Often used in cases where accurate results are preferred over interpretable results

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY *of* WASHINGTON



Relationship with the data science process

Data modeling is a crucial part of the data science process that follows data engineering

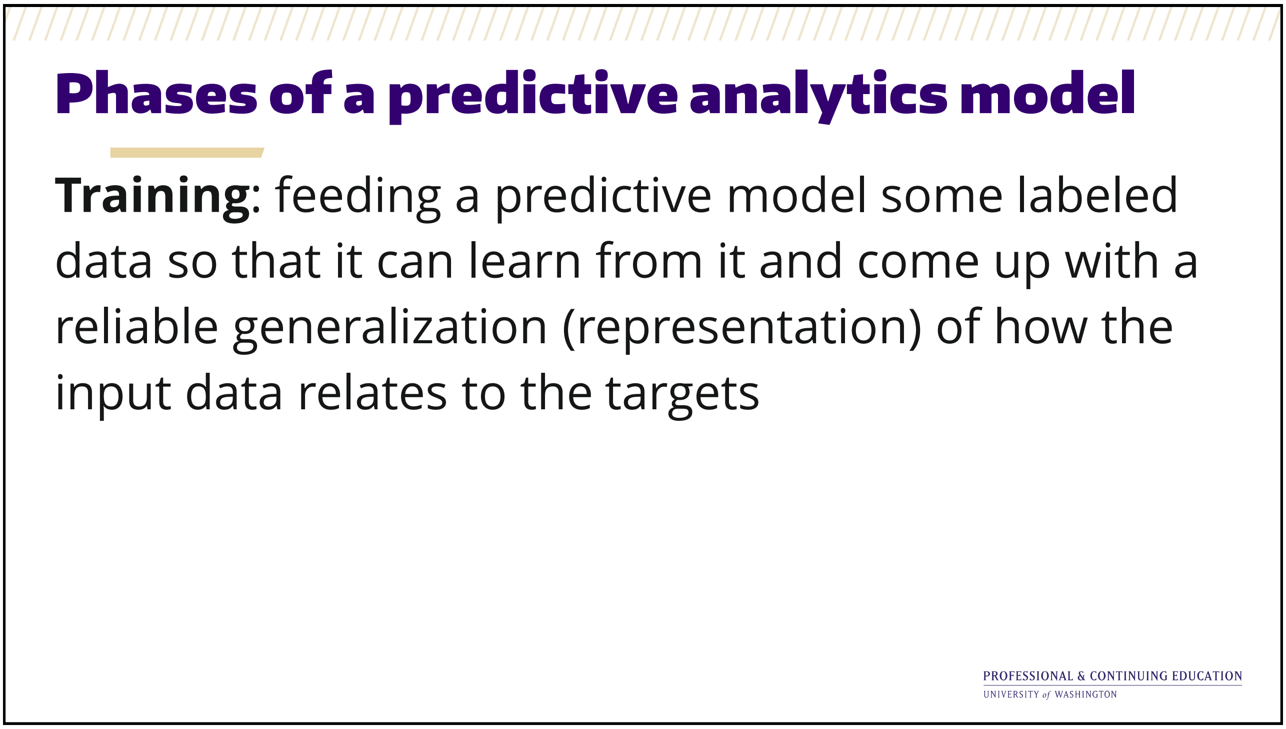
- Not to be done at the expense of other parts though, e.g. feature engineering

Closely linked to data visualization, which usually ensues

Data modeling is usually done natively, though in some cases it is done on the cloud

- Models developed in this stage are used later on as part of a data product

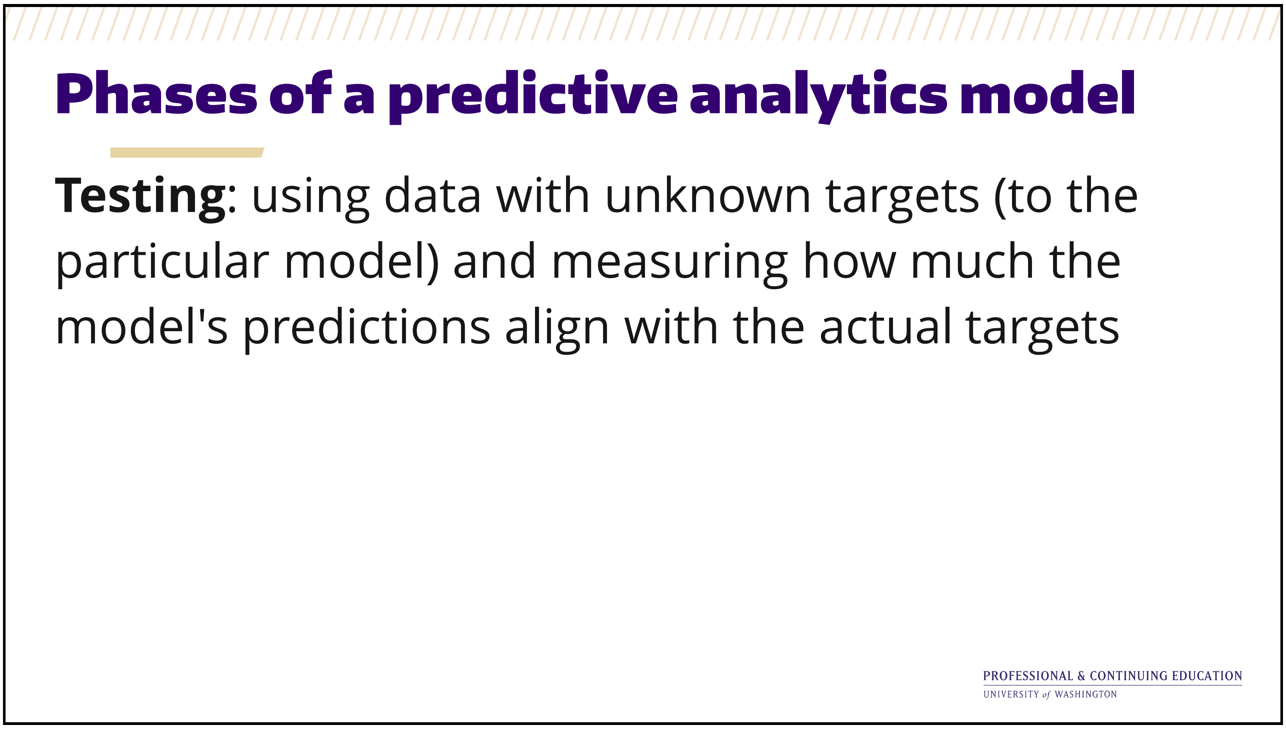
PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY *of* WASHINGTON



Phases of a predictive analytics model

Training: feeding a predictive model some labeled data so that it can learn from it and come up with a reliable generalization (representation) of how the input data relates to the targets

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY *of* WASHINGTON



Phases of a predictive analytics model

Testing: using data with unknown targets (to the particular model) and measuring how much the model's predictions align with the actual targets

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY *of* WASHINGTON

Phases of a predictive analytics model

Deployment: putting a tested predictive model into production, to be used with unknown targets (to both the classifier and to us)

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY *of* WASHINGTON

Classification

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY *of* WASHINGTON

Overview of Classificiation

Classification = a predictive analytics methodology, where the variable predicted (target) is categorical (e.g. animal types in a picture)

– When the target variable is binary → binary classification

Classification is one of the most research methodology in predictive analytics, with numerous methods and techniques for it

– Many data science problems can be framed as classification ones

Examples & Applications

- Classifying whether a patient has a certain disease or not
- Figuring out if someone is going to default a loan or not
- Predicting what someone is going to vote in an election
- Predicting which team is going to win a given league
- Classifying if a customer is going to buy something or not
- Predicting if a stock price is going to rise or drop
- Figuring out if the sentiment of a given text is positive, negative, or neutral

Things to keep in mind

Classification, like every predictive analytics methodology, is not 100% accurate

- The classification output (class) is often accompanied by some metric of certainty, usually in the form of a probability the classification is correct

A classification system (classifier) is reliable if it predicts something accurately and with high certainty

- Classification performance often includes other factors, such as time and resources required

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Data involved in classification models

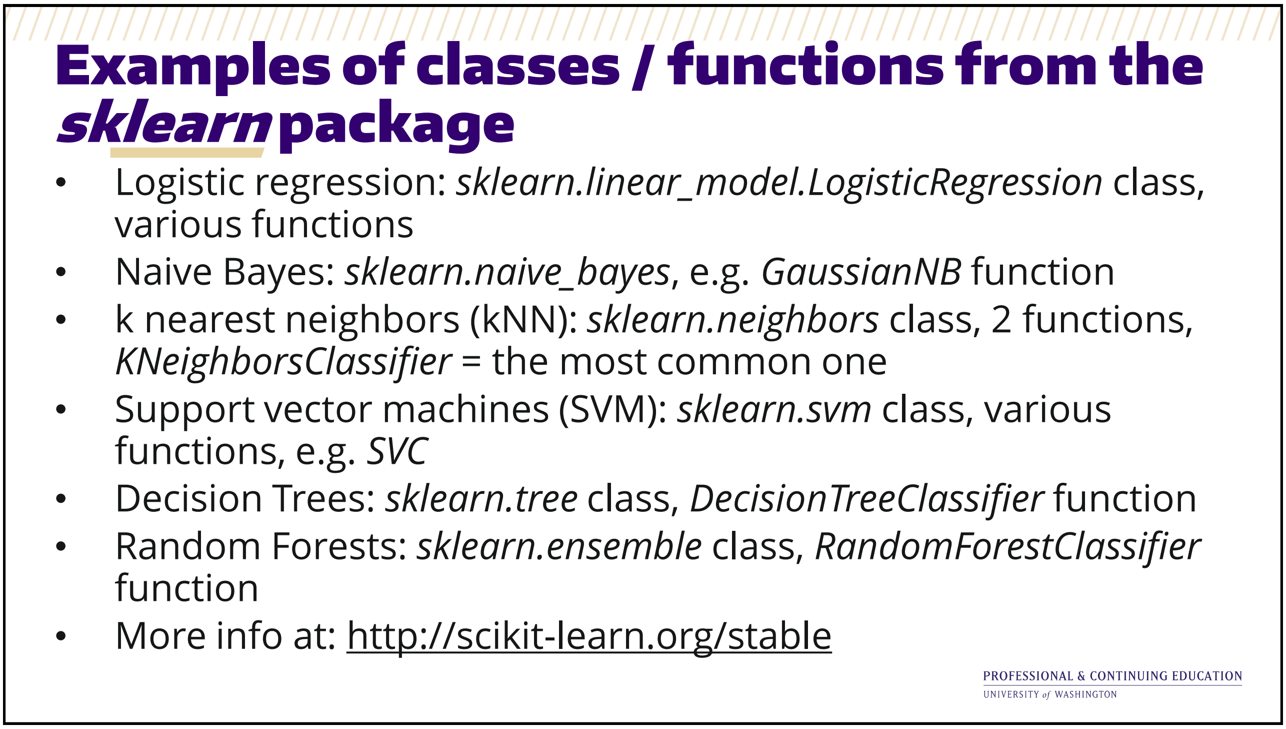
Inputs (features): X (a matrix)

Labels: Y (a vector)

Predictions of classifier: Y_hat, C (both vectors)

Parameters: various, depending on the classifier function

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON



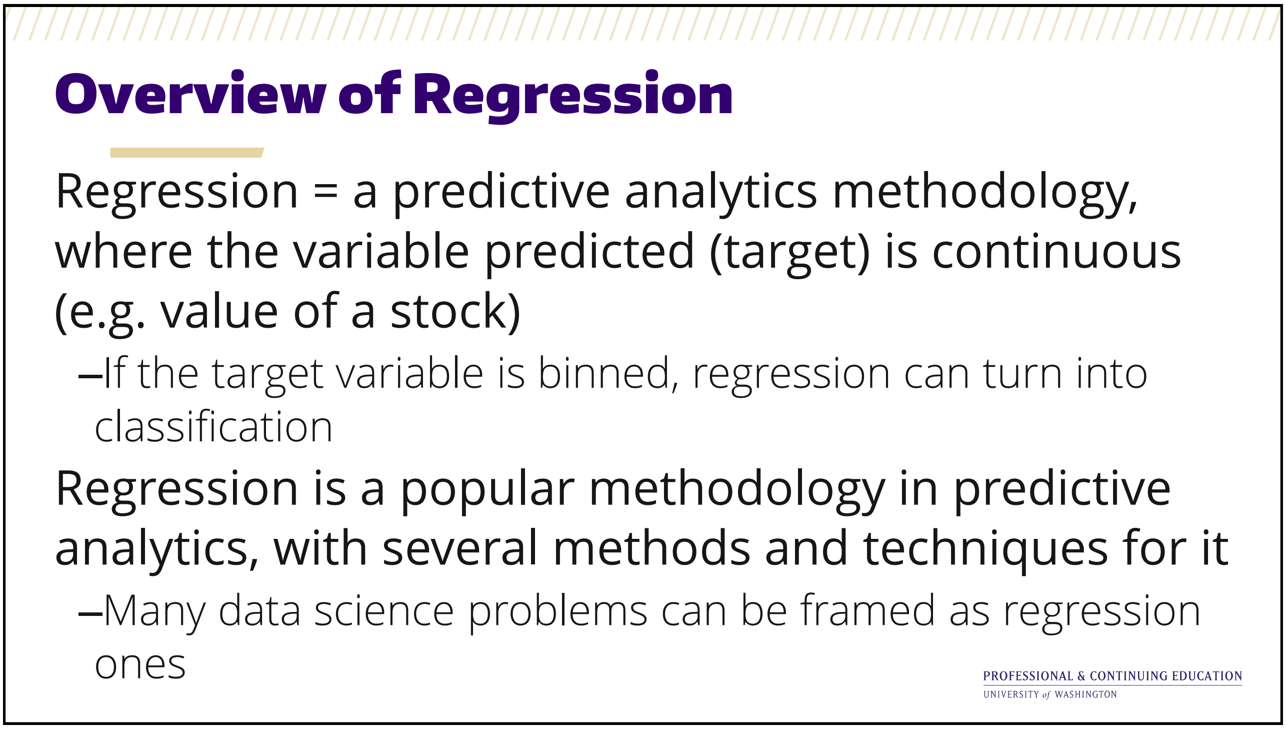
Examples of classes / functions from the *sklearn* package

- Logistic regression: *sklearn.linear_model.LogisticRegression* class, various functions
- Naive Bayes: *sklearn.naive_bayes*, e.g. *GaussianNB* function
- k nearest neighbors (kNN): *sklearn.neighbors* class, 2 functions, *KNeighborsClassifier* = the most common one
- Support vector machines (SVM): *sklearn.svm* class, various functions, e.g. *SVC*
- Decision Trees: *sklearn.tree* class, *DecisionTreeClassifier* function
- Random Forests: *sklearn.ensemble* class, *RandomForestClassifier* function
- More info at: http://scikit-learn.org/stable

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY *of* WASHINGTON

Regression

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY *of* WASHINGTON



Overview of Regression

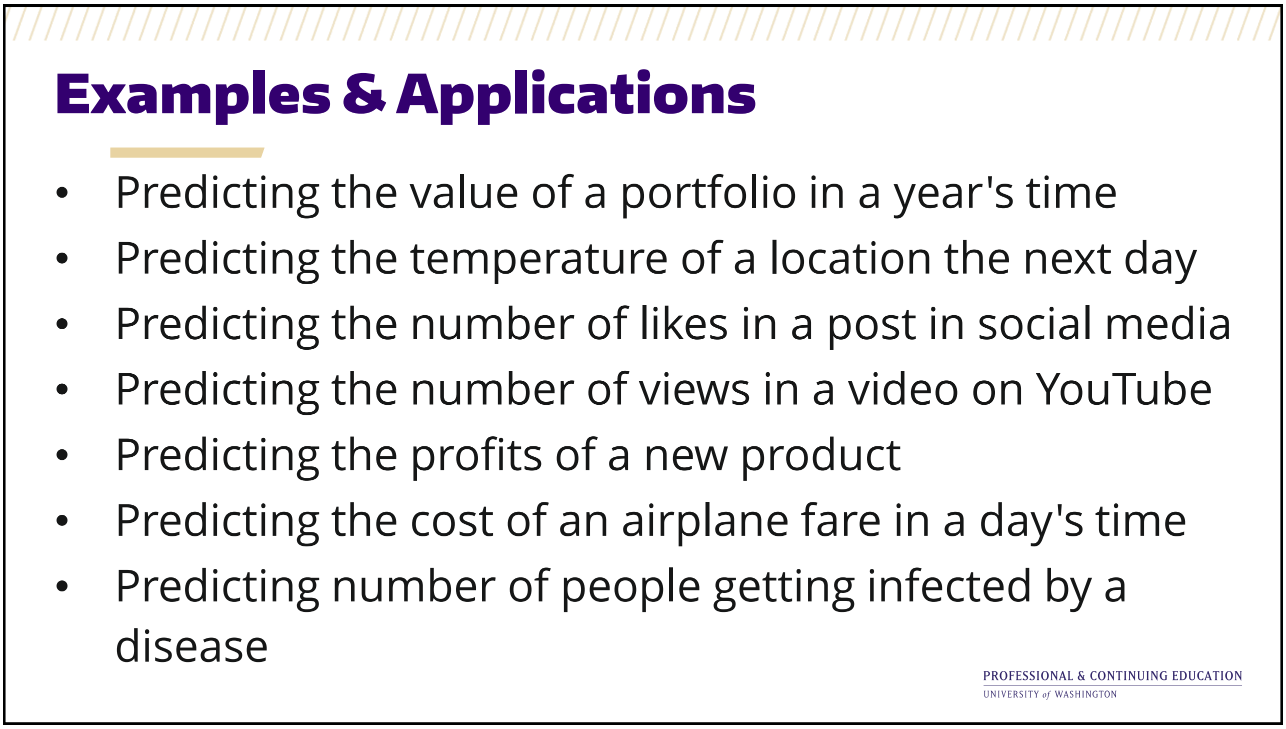
Regression = a predictive analytics methodology, where the variable predicted (target) is continuous (e.g. value of a stock)

- –If the target variable is binned, regression can turn into classification

Regression is a popular methodology in predictive analytics, with several methods and techniques for it

- –Many data science problems can be framed as regression ones

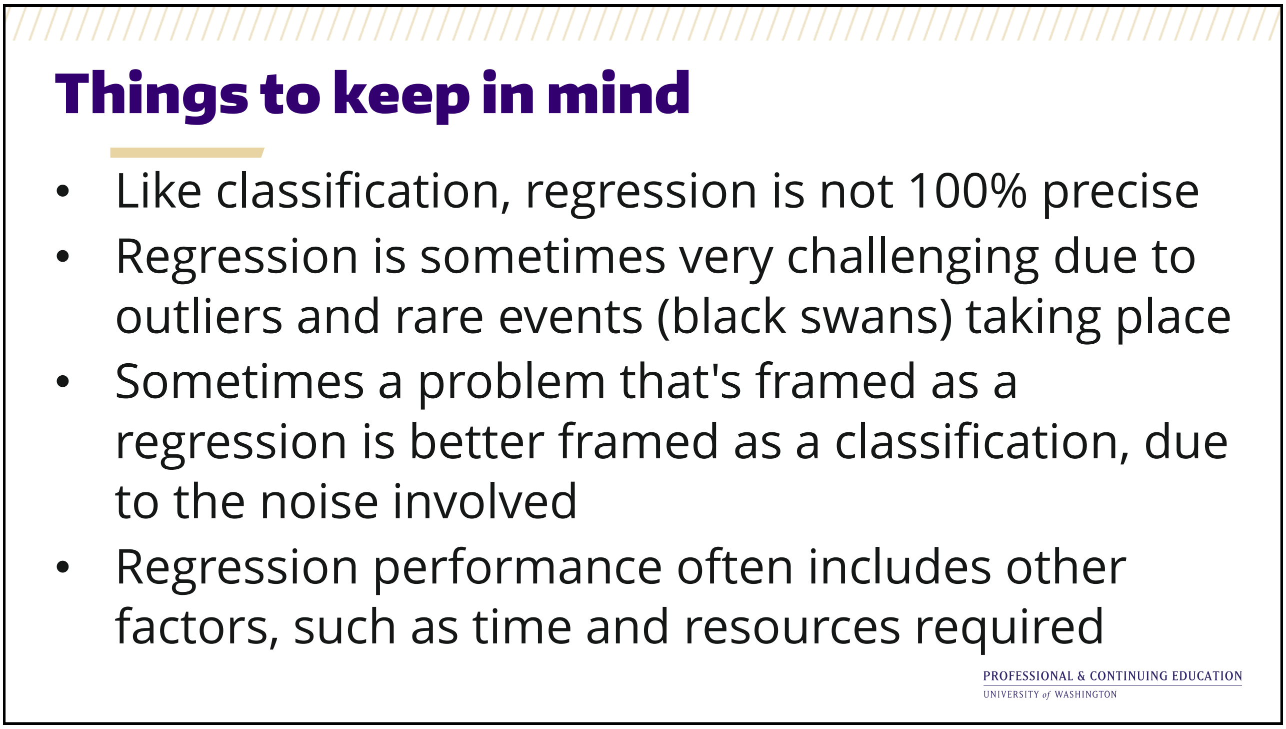
PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY *of* WASHINGTON



Examples & Applications

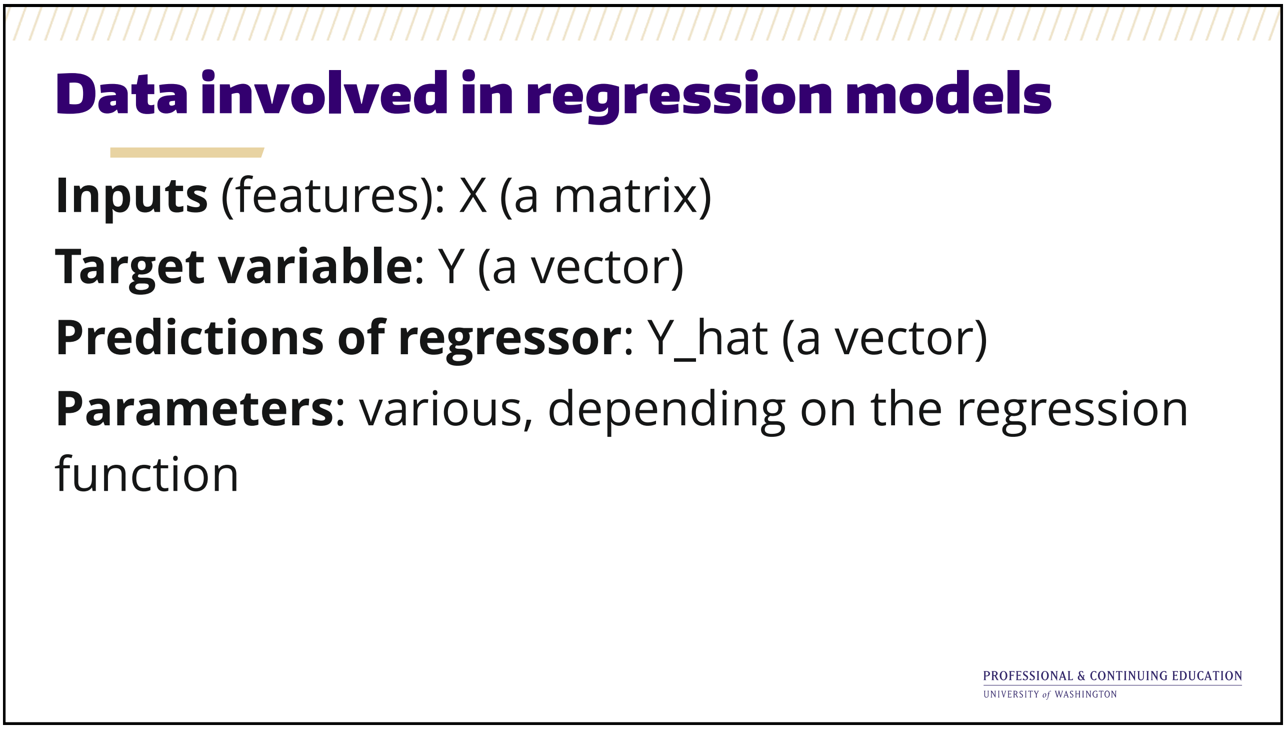
- Predicting the value of a portfolio in a year's time
- Predicting the temperature of a location the next day
- Predicting the number of likes in a post in social media
- Predicting the number of views in a video on YouTube
- Predicting the profits of a new product
- Predicting the cost of an airplane fare in a day's time
- Predicting number of people getting infected by a disease

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY *of* WASHINGTON



Things to keep in mind

- Like classification, regression is not 100% precise
- Regression is sometimes very challenging due to outliers and rare events (black swans) taking place
- Sometimes a problem that's framed as a regression is better framed as a classification, due to the noise involved
- Regression performance often includes other factors, such as time and resources required



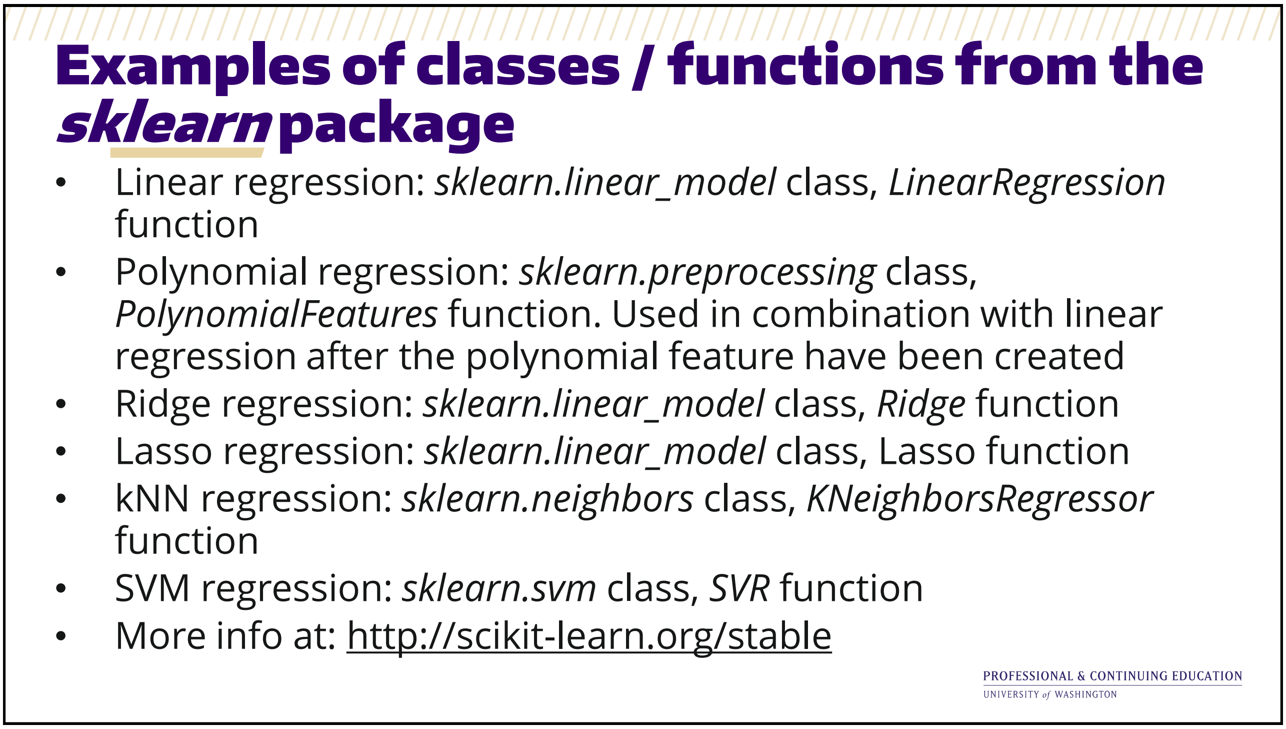
Data involved in regression models

Inputs (features): X (a matrix)

Target variable: Y (a vector)

Predictions of regressor: Y_hat (a vector)

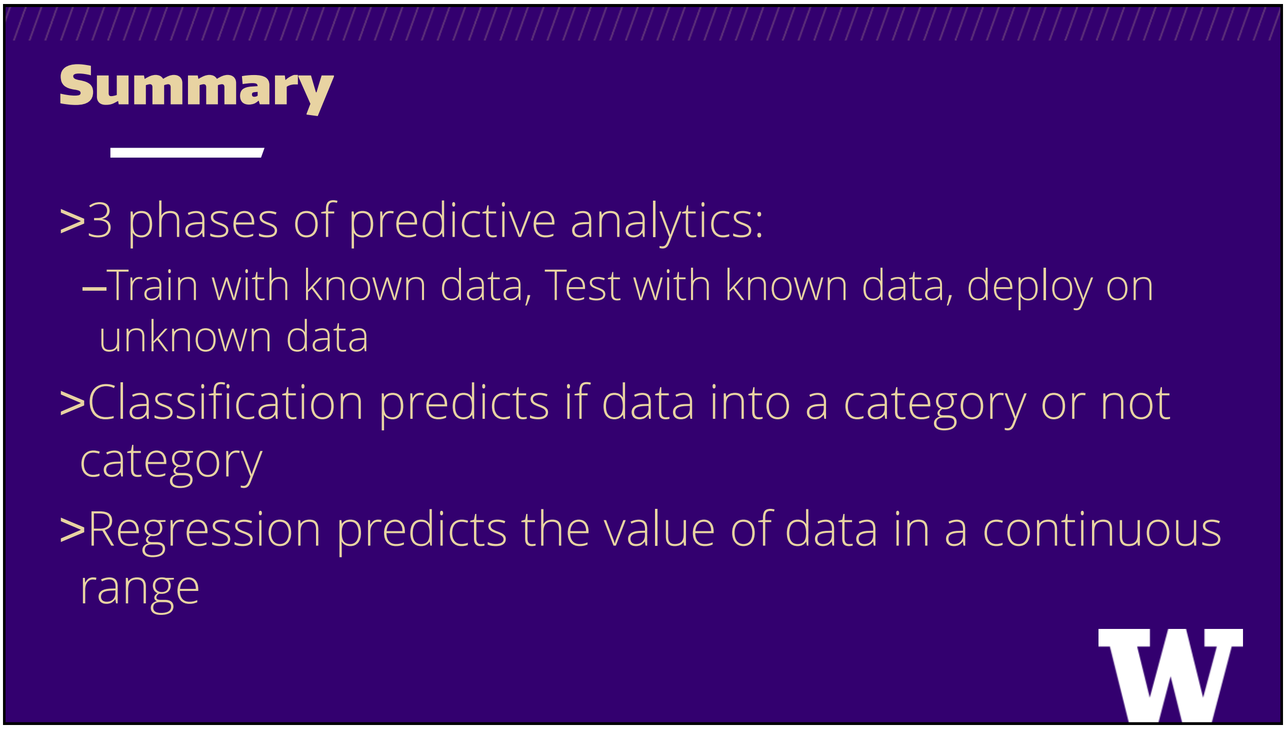
Parameters: various, depending on the regression function



Examples of classes / functions from the *sklearn* package

- Linear regression: *sklearn.linear_model* class, *LinearRegression* function
- Polynomial regression: *sklearn.preprocessing* class, *PolynomialFeatures* function. Used in combination with linear regression after the polynomial feature have been created
- Ridge regression: *sklearn.linear_model* class, *Ridge* function
- Lasso regression: *sklearn.linear_model* class, Lasso function
- kNN regression: *sklearn.neighbors* class, *KNeighborsRegressor* function
- SVM regression: *sklearn.svm* class, *SVR* function
- More info at: http://scikit-learn.org/stable

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON



Summary

>3 phases of predictive analytics:

–Train with known data, Test with known data, deploy on unknown data

>Classification predicts if data into a category or not category

>Regression predicts the value of data in a continuous range