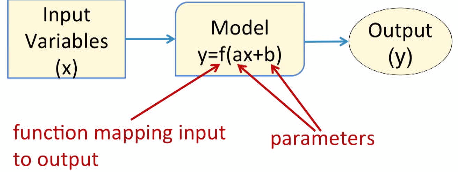
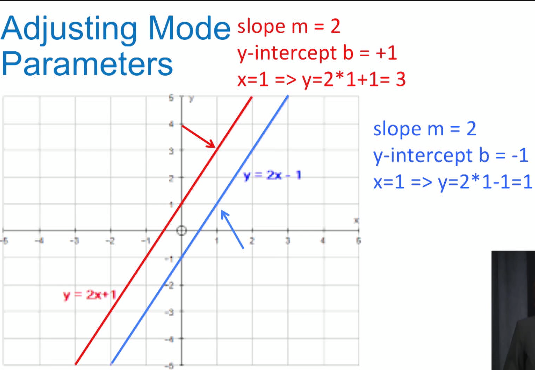
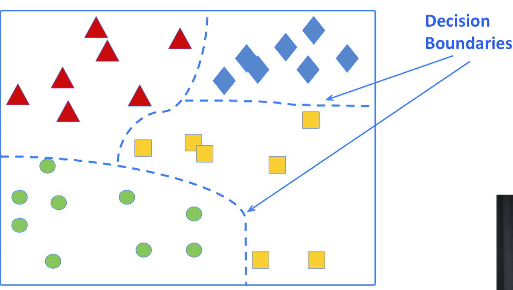
* Machine Learning = field of study focusing construction of CPU systems (**models**) that can *learn from data* w/out being explicitly programmed to
* i.e. learn to perform a specific task by analyzing lots of examples for a particular problem
* ex: recognize image of cat after being shown lots of images of cats
* Algorithms are programmed to learn *from data*, NOT from a set of given instructions
* i.e. determines what features are important in recognizing a cat
* therefore its important note the amount + the quality of data provided to a ML model
* ML models can be used to discover hidden trends + patterns which can lead to valuable insights into data to allow for data-driven decisions to be made for particular problems
* **ML Model =** mathematical mode/**parametric function** that maps inputs to outputs



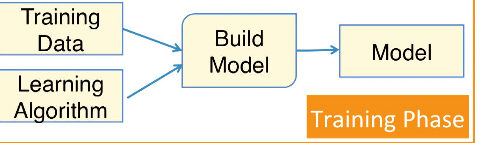
* i.e. model has **input/parameters** + uses equations to determine the relationship between inputs and outputs
* **Parameters** are used by the model to modify inputs the generate output
* They are adjusted to correct/refine the input-output relationship
* i.e. adjusting for **a** and **b** above will adjust how **x** maps to **y**
* given x, a model uses its parameters (a + b) to determine y

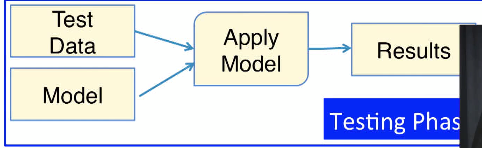


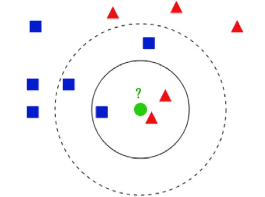
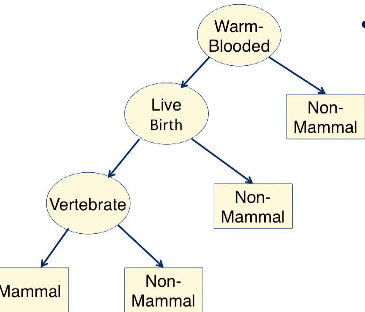
* Parameters of ML models are adjusted/estimated from the data using a learning algorithm
* The above process = **model building, models creation, model training, model fitting, etc.**
* Model parameters are adjusted in order to *reduce error* (getting outputs to match the desired outputs/targets as much as possible in classification or regression)
* Applications:
* Credit card fraud detection
* Each transaction is analyzed against transaction history to determine if its legitimate or not
* Suspicious activity 🡪 big ticket items in a category you’ve never shown interest in, Point of sales in foreign countries
* If suspicious, transaction could be denied or you may be asked to confirm the purchase
* Handwritten digit recognition 🡪 read #’s on checks to determine bank accounts, deposit amount, etc.
* ML models can sift through different variations of digits to find similar patterns to distinguished 1 from 9
* Website recommendations 🡪 list of related items after purchases
* **Data mining** (process of finding patterns in databases + data warehouses) is similar to ML + they both often use the similar algorithms and techniques
* **Predictive** **Analytics** is also similar 🡪 analyze data to predict future outcomes (sales forecasting, customer behavior)
* **Classification** 🡪 predict category/label of input data
* Examples:
* Weather as sunny, rainy, windy, cloudy
* Input data = sensor data specifying temperature, relative humidity, atmospheric pressure, wind speed + direction, etc.
* Target = the categories
* Tumor as benign or malignant (**binary classification**)
* ID handwritten digits as 1 of 10 categories (0-9)
* Given specific values for different features, the model is tasked w/ coming up w/ the category/classification/label of the record
* Output = target/label/outcome/class variable/class/category
* Target is ALWAYS categorical
* Can be **binary** (only 2 possible classifications) or **multi-class** (multinomial/multi-label)
* Think of classification as carving up an input space into regions corresponding to different labels

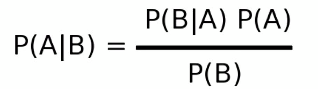


* The model formed the **decision boundaries** used to define regions separating the shapes
* Classifications are based on these regions
* To build a classification model/**classifier**, we need to use data to adjust model parameters to form decision boundaries to separate target classes
* 2 phases of building a classifier (+ other ML models)
* 1) **Training** 🡪 model is constructed + parameters are adjusted using training data
* 2) **Testing** 🡪 learned model is applied to *new* data, never-before-seen (not in training)





* Model is then evaluated on how it performs to test data
* **Goal** = have classifier perform well on both training + testing data = get model outputs to match targets as much as possible
* To do this, a classifier adjusts is parameters by applying a learning algorithm in the training phase
* Models used to build a classifier:
* **k-nearest neighbor (kNN) -** relies on the notion that samples w/ similar values for inputs *likely* belong to the same class
* 
* i.e. classification is dependent on target values of neighboring points
* **Decision tree** 🡪 uses a tree-like structure to represent multiple decision paths
* Each path leads to a different way to classify an input sample
* **Random forest 🡪** multiple decision trees
* 
* **Naïve Bayes** 🡪 uses probabilistic approach to classification
* Baye’s theorem is used to capture the relationship between input data + output class
* *compares probability of an event in the presence of another event*



* **Regression** 🡪 predicting numeric value
* Ex: predict stock price in $
* Predicting stock prices as “rise” or “fall” 🡪 classification
* Ex: Predict *amount* of rain in a region
* Predicting “rain” or “not rain” tomorrow 🡪 classification
* **Cluster** 🡪 group similar items in a dataset into groups
* Ex: customer segmentation
* seniors, adults, teens 🡪 different likes/dislikes + purchase behavior
* can provide different targeted marketing to each
* **Association analysis 🡪** come up w/ a set of rules to capture associations between items/events
* Rules are used to determine when they occur together
* Ex: Market Basket analysis 🡪 used to understand customer purchase behavior
* Ex: Revealing that customers w/ checking accounts tend to be interested in other investment vehicles, such as money market accounts + this info can be used for cross-selling
* i.e. selling money market accounts to customers w/ CD’s
* Ex: Amazon recommender systems based on items usually purchased together
* Ex: Identification of web pages often accessed together 🡪 can generate more offers on related web pages
* **Supervised learning** 🡪 target/outcome is provided (**labelled data**)
* Classification + regression, generally
* **Unsupervised learning** 🡪 target/outcome is NOT known
* Clustering and association analysis, generally
* **Sample/Record/Observation =** instance/example of an **entity** of the data (i.e. a row in a dataset)
* **Variables/Features/Attribute/Dimension/Field** = different info pieces about samples
* Each has a data type associated w/ it 🡪 numerical, categorical, string, binary, date, etc.
* Numeric/Quantitative
* Can be measured and/or sorted
* Can be continuous or integers, and be positive, negative, or both
* Categorical/qualitative/Nominal
* Variable w/ labels, names, categories, etc. that describe characteristics of an entity
* **Scitkit-learn 🡪** ML library in Python
* Open-source + built on top of Numpy, SciPy, matplotlib
* Rapidly developed + improved by an active community of developers (800+ contributors)
* Includes end-to-end ML 🡪 entire data science process (i.e. data cleaning, transformations, + ML)
* Supports whole data science process via functions for transformations, cleaning, prep, scaling, normalization, feature engineering, + missing value handling
* **Utility functions** for transforming raw features to suitable format
* **API** for scaling (remove mean + keep unit variance), normalization to have a unit norm, **binarization**, **One Hot Encoding** for categorical features, generating **high-order features**, building custom transforms, handling missing values
* Also has built-in functions for many ML algorithms, ready for modeling + analysis
* Still requires some expertise to use these for the right tasks, but many online resources smooth out the learning curve
* Documentation includes tutorials
* Has specialized implementations for **dimensionality reduction algorithms**
* Enables us to reduce features while preserving variance
* **Principal component analysis (PCA), single value decomposition, factor analysis, independent component analysis, matrix factorization, latent dirichlet allocation**
* Has methods for model selection
* **Cross-validation** methods
* Library functions for **tuning hyper parameters**
* **Model evaluation mechanisms** to measure model performance
* Plotting methods for visualizing scores to evaluate models