Introduction to Machine Learning

Lisa Orr and Ozlem Senlik

Quick Questionnaire

How many people comfortable/familiar with basic …?

1. Linear algebra: Matrix multiplications
2. Probability and statistics: Distributions, Bayes Theorem
3. Calculus

- These are not prerequisites for this workshop. No worries..

Outline for full day workshop

- What is machine learning? 9:30-10:15
  - Types of learning 30 min
  - Lab1 15 min

BREAK(15 min)

- Machine learning techniques I 10:30-11:10
  - Introduction of machine learning techniques 10 min
  - Training concepts
  - Brief inner workings of linear regression
    - Model representation
    - Cost function
    - Model parameter initialization
    - Model tuning-gradient descent
- Machine learning techniques II 11:10-12:00
  - Inner workings of logistic regression
    - Model representation
    - Cost function
    - Probabilistic interpretation
  - Regularization
  - Lab2 (15 min)

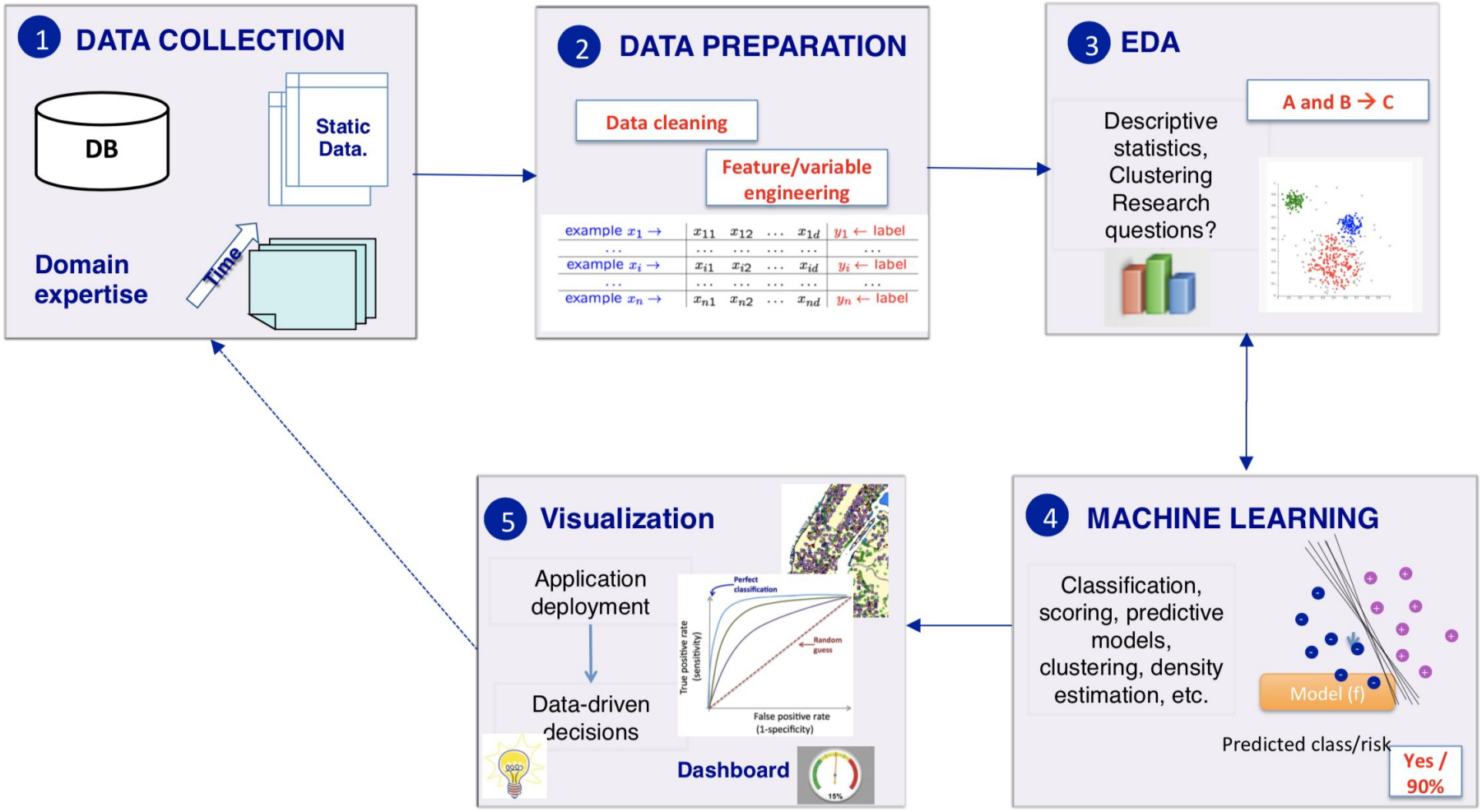
LUNCH 12-1pm

1:-1:30: Getting Into ML and Data Science

- Evaluation and Model Selection 1:30-2:30
  - Generalization
  - Bias Variance Trade off
  - Train/Val/Test Splitting
  - Train/Validation/Test Distributions
  - Lab3 (20 min)
- Break 2:30-2:45
- Lab 4 2:45-4 pm

Lecture 1: What Is Machine Learning?
9:30-10:15

The Data Science Process



Data Types

Data comes in different types and sizes.

- Text
- Numbers
- Clickstreams
- Graphs
- Tables
- Images
- Transactions
- Videos
- Some or all of above

What is machine learning?

"**Learning is any process by which a system improves performance from experience.**"

Herbert Simon

Machine Learning is the study of algorithms that

- improve their performance ***P***
- at some task ***T***
- with experience ***E***.
- A well-defined learning problem is given by <***P, T, E***>.

Tom Mitchell

Defining the Learning Problem

Improve on task T, with respect to performance metric P, based on experience E

T: Playing checkers
P: Percentage of games won against an arbitrary opponent
E: Playing practice games against itself

T: Recognizing hand-written words
P: Percentage of words correctly classified
E: Database of human-labeled images of handwritten words

T: Driving on four-lane highways using vision sensors
P: Average distance traveled before a human-judged error
E: A sequence of images and steering commands recorded while observing a human driver.

T: Categorize email messages as spam or legitimate.
P: Percentage of email messages correctly classified.
E: Database of emails, some with human-given labels

Slide Credit: Ray Mooney

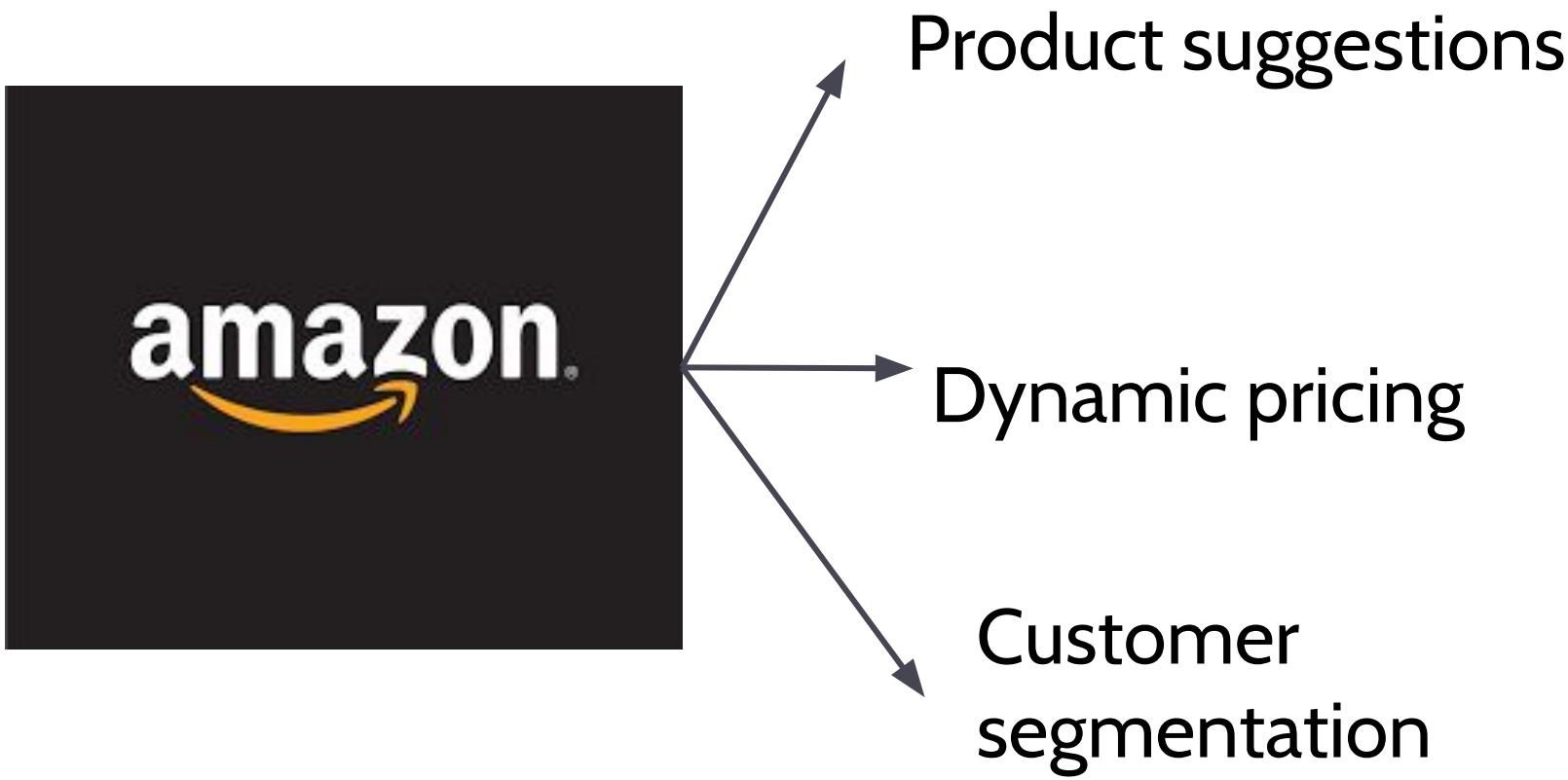
Machine Learning When?

- Human expertise does not exist (navigating on Mars)
- Humans can’t explain their expertise (speech recognition)
- Models must be customized (personalized medicine)
- Models are based on huge amounts of data (genomics)

Why now?

- Flood of available data (especially with IOT)
- Increasing computational power
- Growing progress in available algorithms and theory developed by researchers
- Increasing support from industries

ML in Our Lives



ML in Our Lives

Suggest pick up/drop off locations

Uber share finds customers travelling in similar routes

Allocate bookings based on time, distance, traffic, price and ratings

What machine learning is not?

- Machine learning is not an exact science
  - No ML method works with 100% precision.

- Precision actually means chances of success.
  - 90% precision means that the algorithm is precise in 90% cases.

Is it bad that machine learning is not exact?

Programmer $100 → exact interest → *Programmer* $100.000001

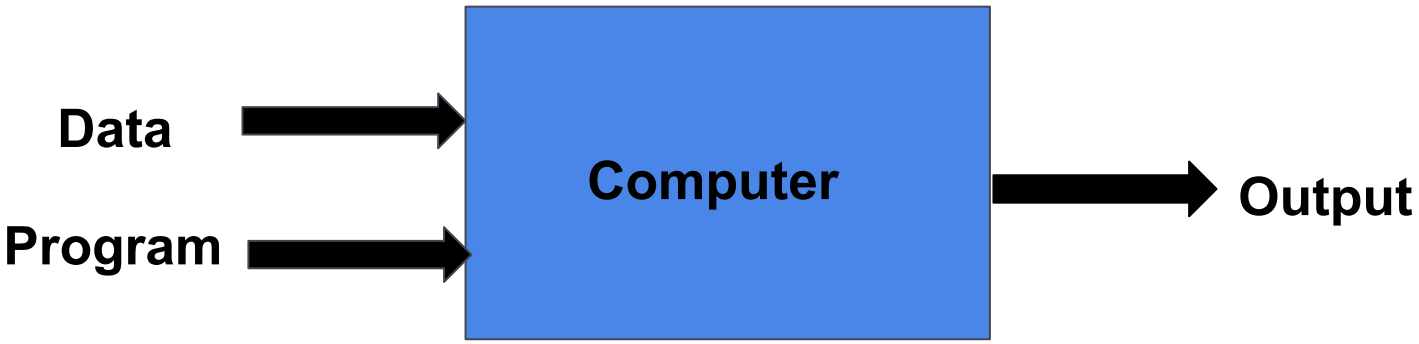
Bank $0 → ML calculated chances → *Bank* $100

Terminology

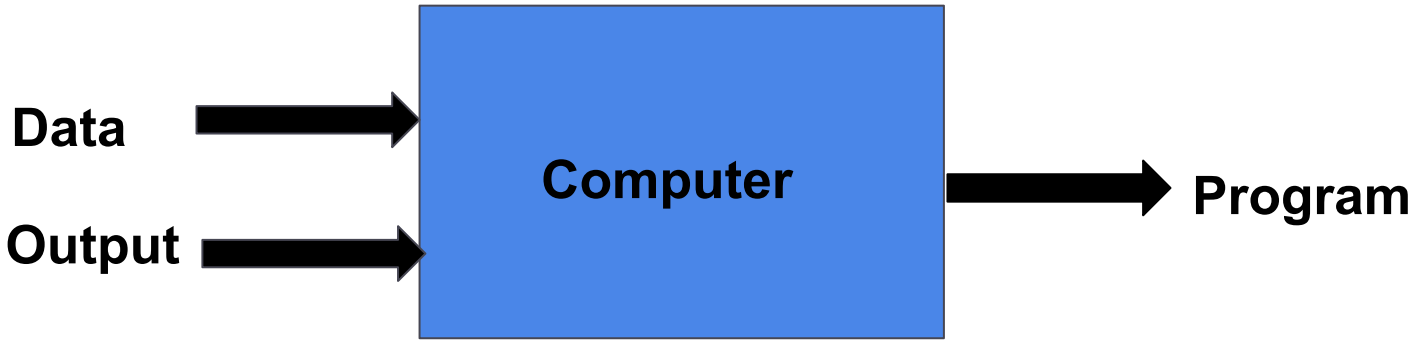
- Features
  - Distinct traits that can be used to describe each item/observation quantitatively.
- Samples
  - A sample is an item in the data set to process (e.g. classify). It can be a document, an image, a video, a row in a database or csv file, or whatever you can describe with a fixed set of quantitative traits.
- Feature extraction
  - Preparation of numeric features from given raw data.
  - Transforms the data in the high dimensional space to a space of fewer dimensions.
- Training Set
  - Set of data to discover potentially predictive relationships.

Traditional Programming vs Machine Learning

Traditional Programming: Following Instructions

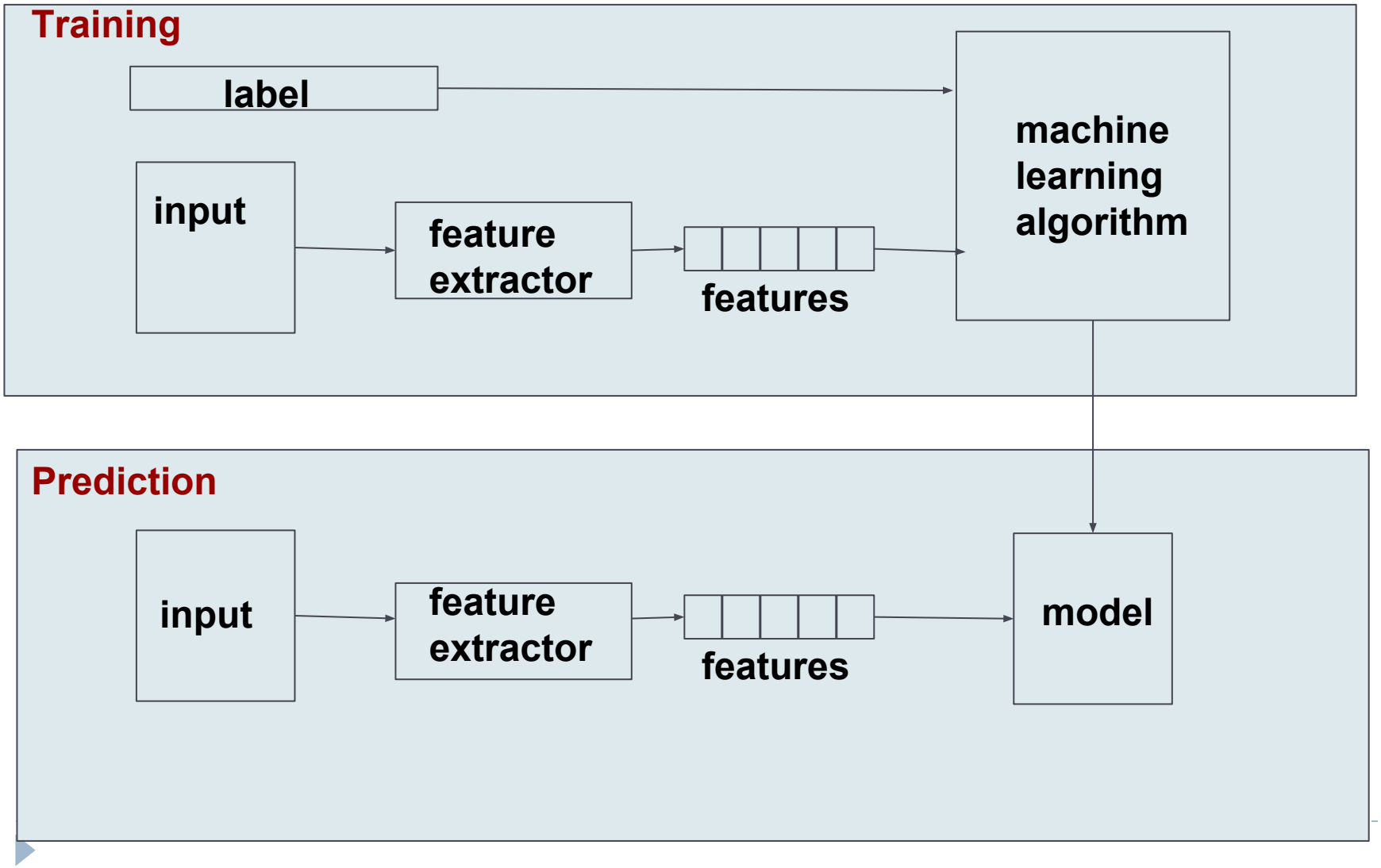


Machine Learning



Slide Credit: Pedro Domingos

Workflow in a Nutshell



Types of Learning

Supervised (inductive) learning

– Given: training data + desired outputs (labels)

Unsupervised learning

– Given: training data (without desired outputs)

Semi-supervised learning

– Given: training data + a few desired outputs

Reinforcement learning

– Rewards from sequence of actions

Supervised Learning

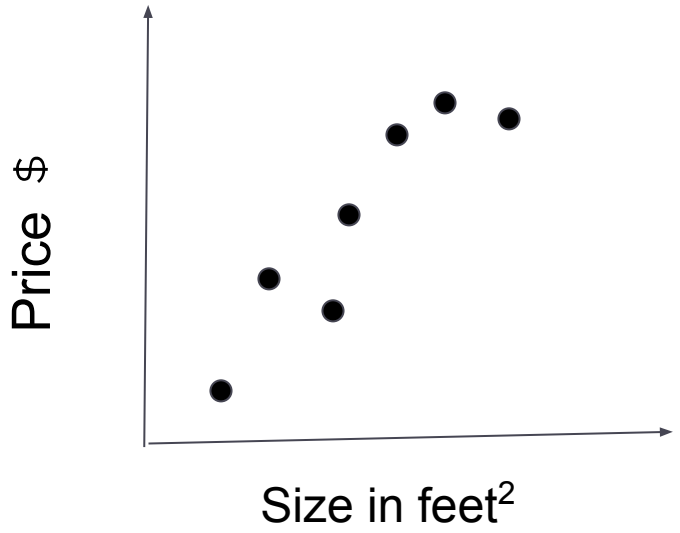
Definition:

In supervised learning, we are given a data set and already know what our correct output should look like, having the idea that there is a relationship between the input and the output.

Example I : Predicting House Prices

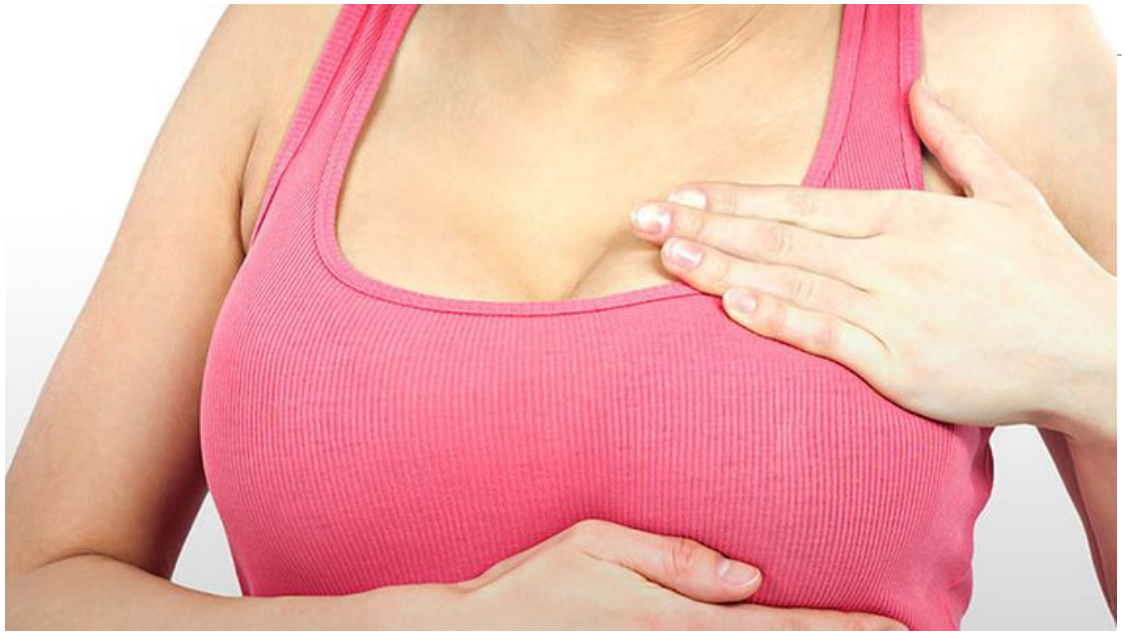
Feature: House size

Correct label: Sold price



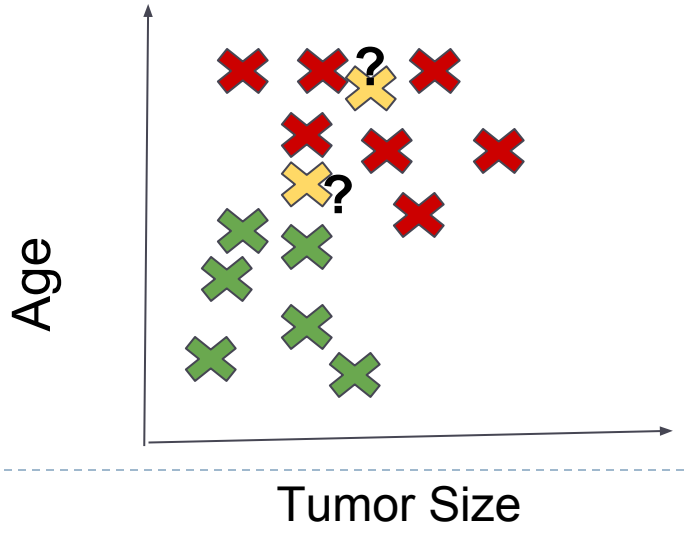
What other features would you use for house price prediction?

Example II : Predicting Breast Cancer



Features: Tumor size, age

Correct label: Tumor type



Unsupervised Learning

Definition

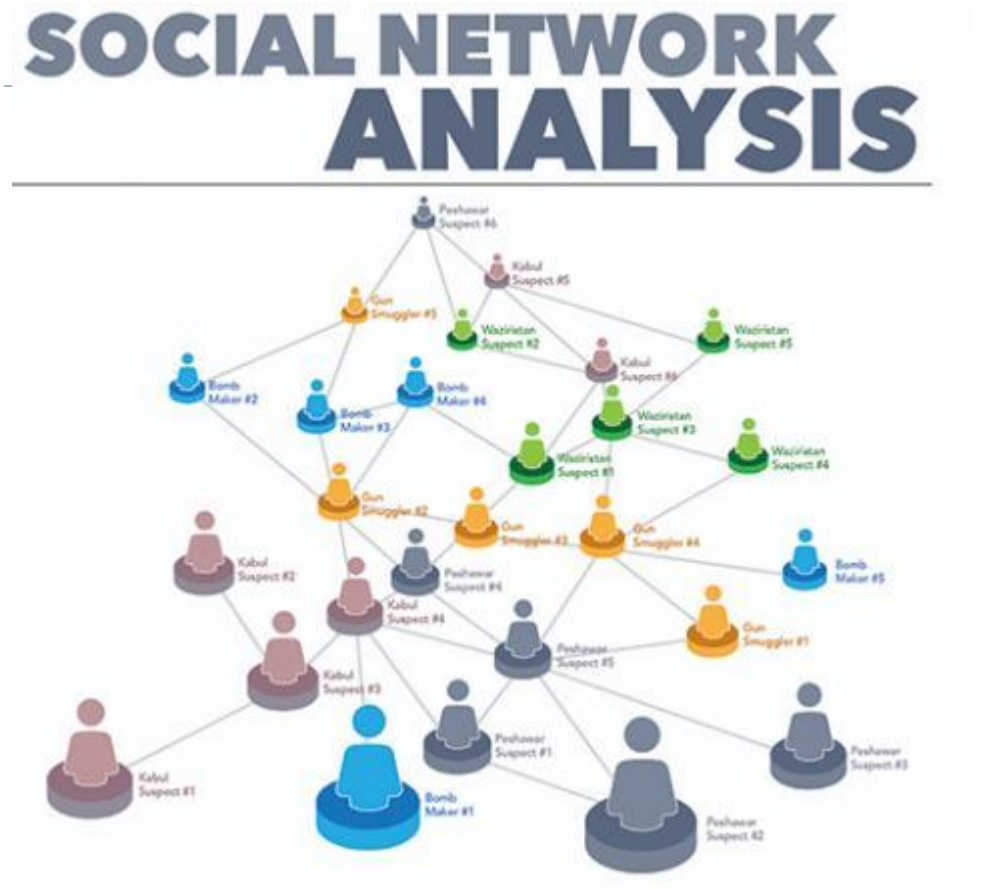
Unsupervised learning allows us to approach problems with little or no idea what our results should look like. We can derive structure from data where we don't necessarily know the effect of the variables.

We can derive this structure by clustering the data based on relationships/similarities among the features in the data.

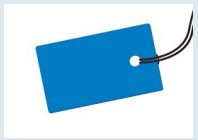
With unsupervised learning there is no feedback based on the prediction results.

Examples

Market Segmentation

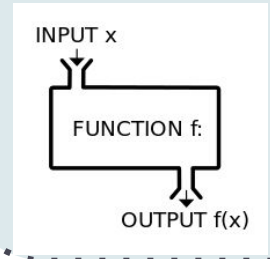


Summary: Supervised vs Unsupervised



Labeled data

Direct feedback



Predict output

Supervised Learning

vs

Unsupervised Learning

Unabeled data

No feedback

Find hidden patterns in data

Exercise I

Please complete in 3 minutes.

Which learning algorithms (supervised or unsupervised) would you use for the following examples?

1. Given a set of emails labeled as spam or not spam, learning a spam filter.
2. Given a set of articles, group them into subset of articles with the same topic.
3. Given demographic and health related information for a set of people, predicting life expectancy.

Exercise II

Please complete in 7 minutes.

Please work in groups to describe a real life problem that can be solved by

1. Supervised learning

2. Unsupervised learning

Please define the task *T*, experience *E* and performance metric *P* for both cases.

LAB I

Lecture 2: Machine Learning Techniques I

10:30-11:10

Machine Learning Techniques Overview

	Supervised Learning	Unsupervised Learning
Discrete	classification categorization	clustering
Continuous	regression	dimensionality reduction

Regression

Attempts to estimate the mapping function (f) from the input variables (x) to numerical or continuous output variables (y).

Let's revisit the example of predicting house prices.

X = ?
Y = ?
f(x) = ?
<P, T, E> for this learning task?

Classification

Attempts to estimate the mapping function (f) from the input variables (x) to discrete or categorical output variables (y).

Let's revisit the example of predicting breast cancer prediction.

X = ?
Y = ?
f(x) = ?
<P, T, E> for this learning task?

Clustering

The process of organizing objects into groups whose members are similar in some way". A *cluster* is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters.

Let's revisit the example of customer segmentation.

Types of Customer Segmentation

1. By location
2. By demographic
3. By benefits
4. By behaviour

- Why segmentation? Ex. Offer targeted deals

Algorithms

Scalability, number of features, outliers, imbalanced classes, interpretability, run time ..

Regression

1. Linear Regression
2. Ridge Regression
3. Support Vector Machines Regression
4. Decision Trees Regression
5. Boosted Trees Regression
6. Random Forest Regression
7. Neural Networks Regression
8. Nearest Neighbors Regression

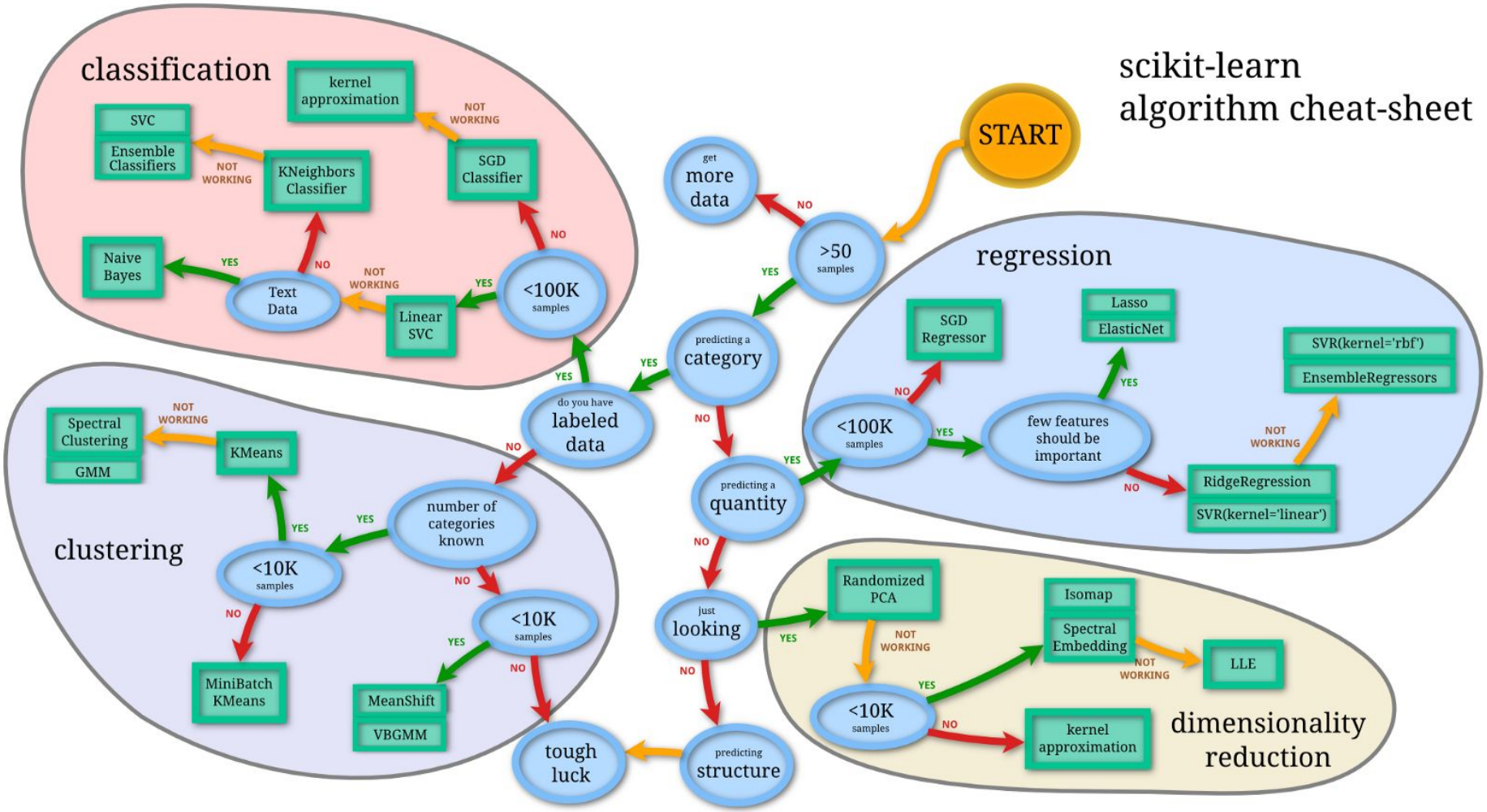
Classification

1. Logistic Regression
2. Naive Bayes Classifier
3. Support Vector Machines
4. Decision Trees
5. Boosted Trees
6. Random Forest
7. Neural Networks
8. Nearest Neighbor

Clustering

1. K-Means
2. Gaussian Mixtures
3. Agglomerative clustering

Scikit-learn guide to selecting algorithms



Exercise III

Please complete in 5 minutes.

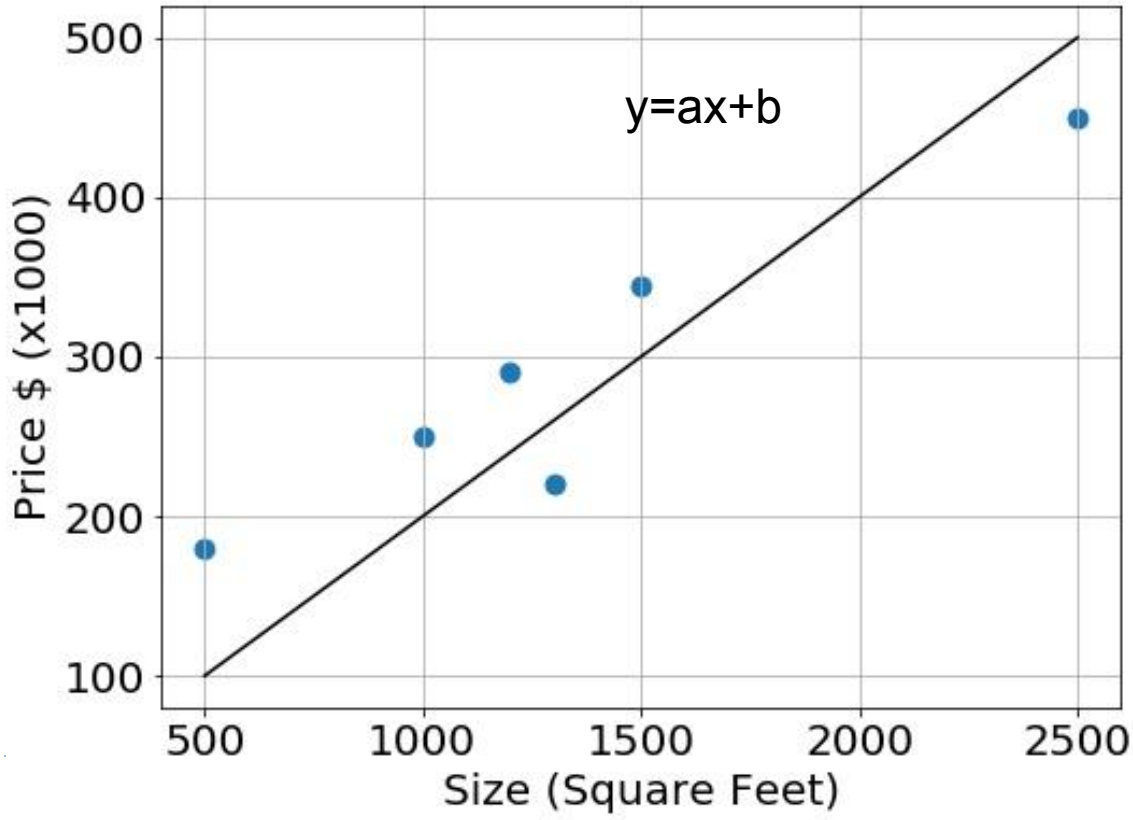
Please work in groups to describe a real life problem that can be solved by

1. Regression
2. Classification
3. Clustering

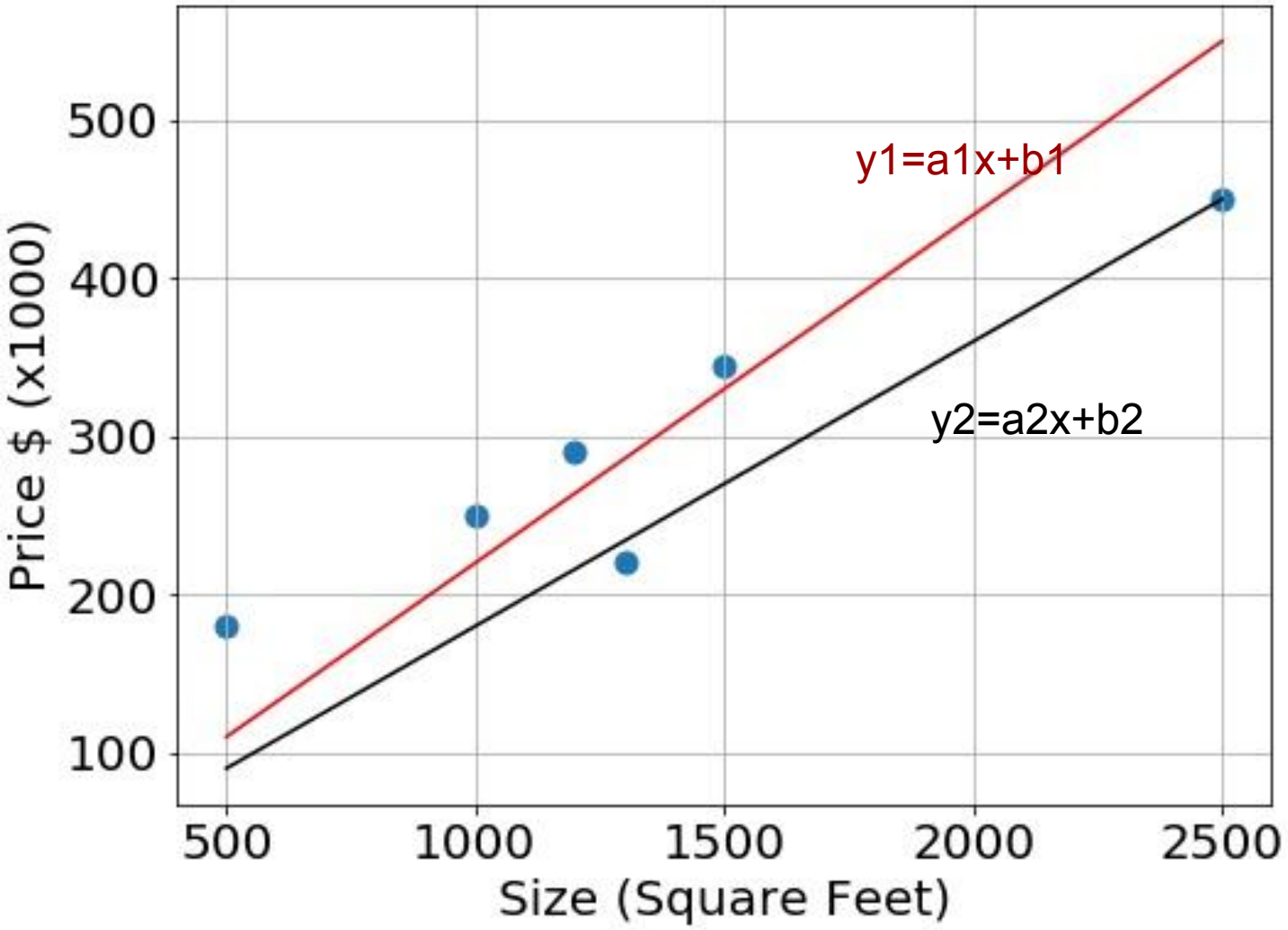
Linear Regression

Model Representation

Size(Feet^2)	500	100	1200	1300	1500	2500	1800
Price $(x 1000)	180	250	290	220	345	450	x
Prediction, y							?

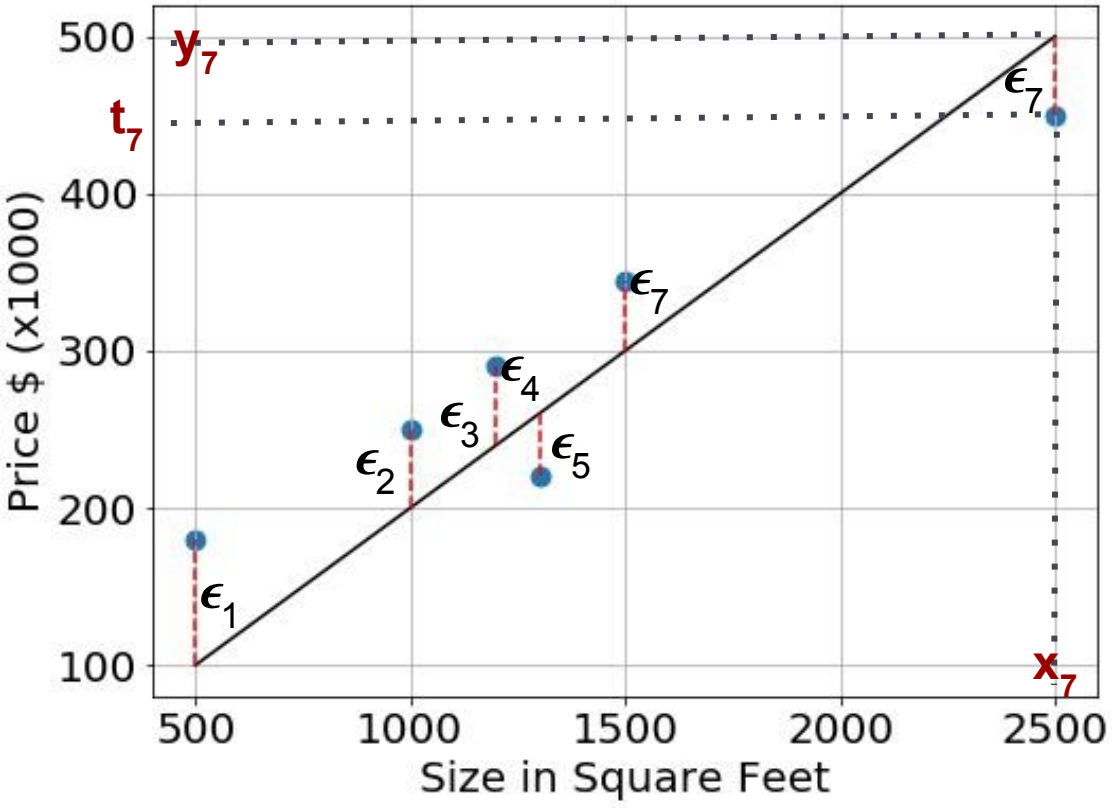


Best model?


y1=a1x+b1
y2=a2x+b2
Price $ (x1000)
Size (Square Feet)
100
200
300
400
500
500
1000
1500
2000
2500

Loss Function

A loss function Loss(x,y,w=(a,b)) quantifies how unhappy you would be if you used w=(a,b) to make a prediction on x when the correct output is y. It is the object we want to minimize.

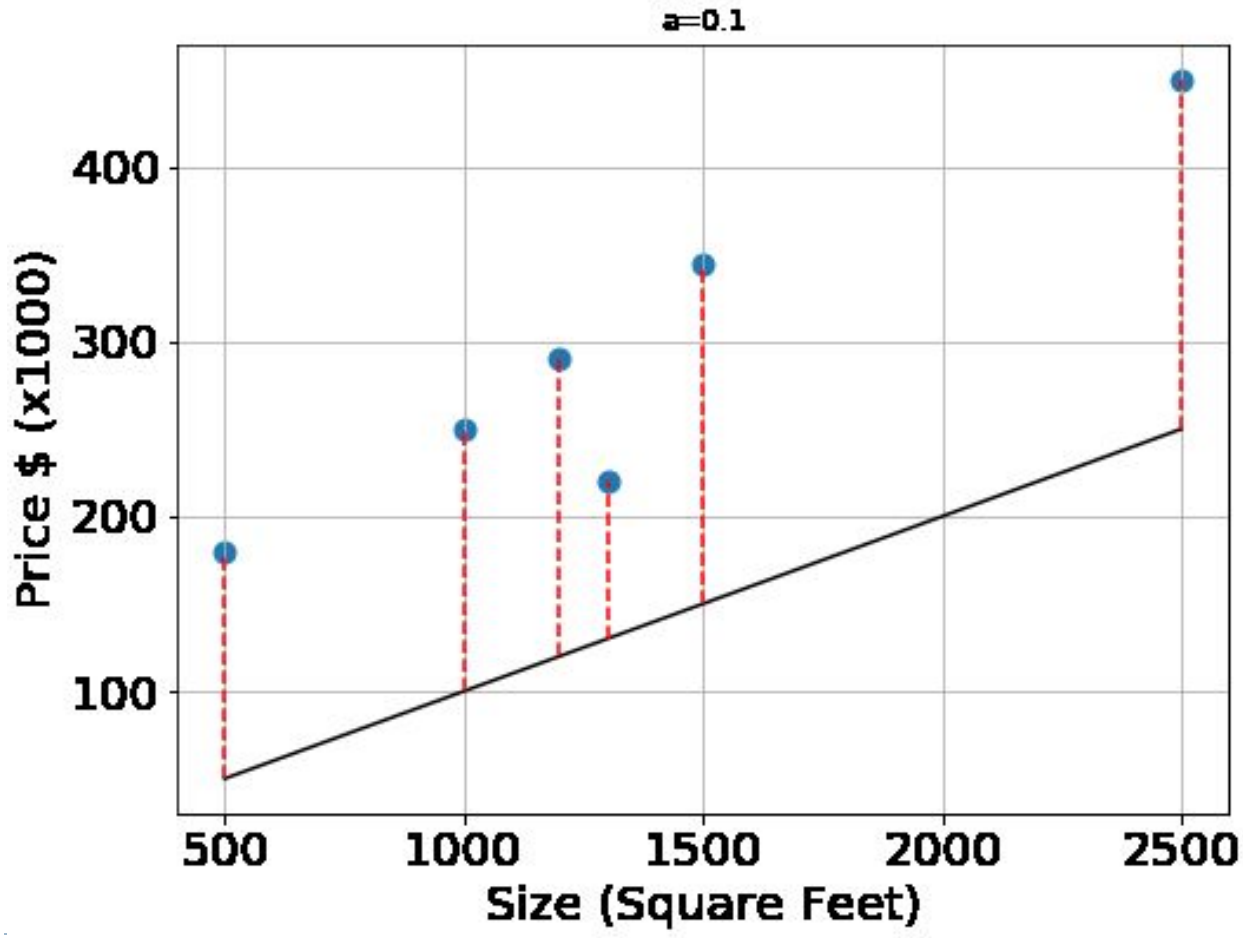


Residuals:

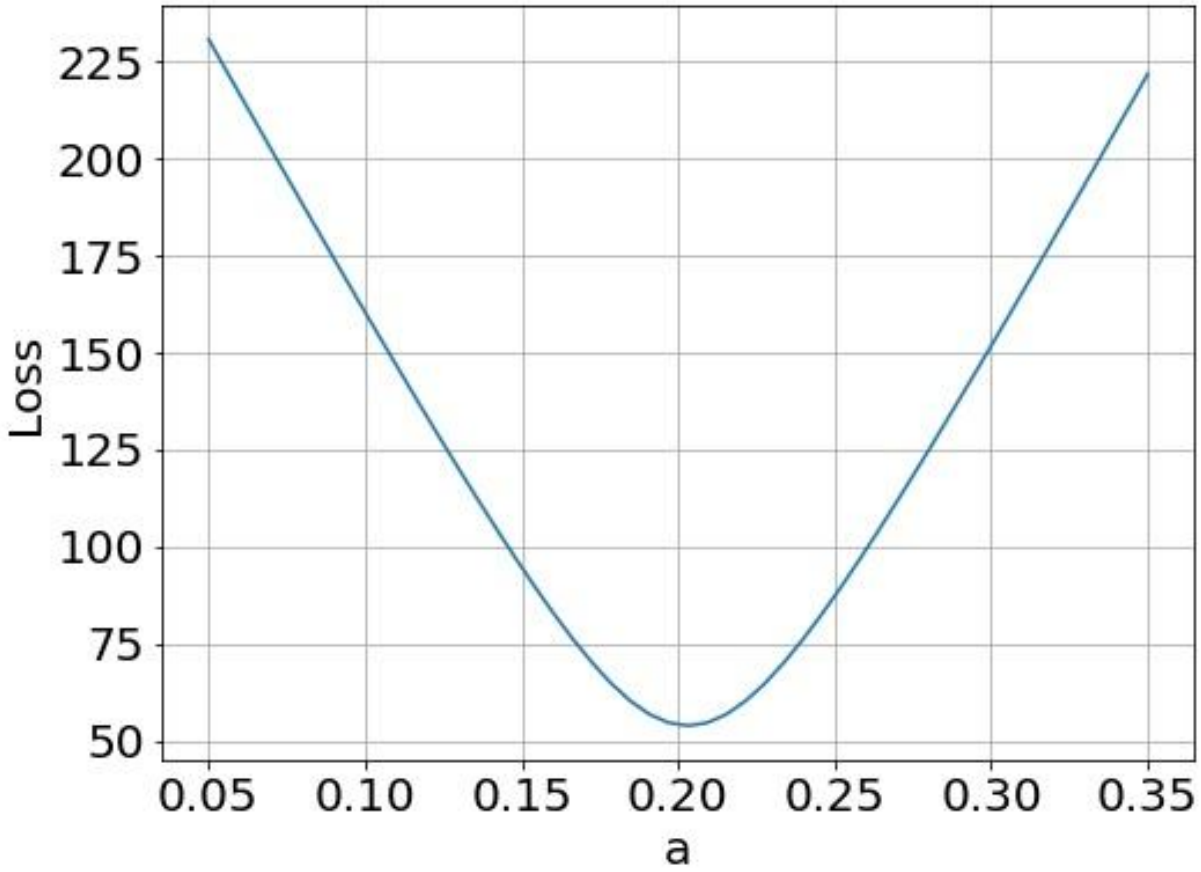
$$\epsilon_7 = y_7 - t_7$$
$$\epsilon_7 = ax_7 + b - t_7$$

$$\text{Loss} = (\epsilon_1^2+\epsilon_2^2+\epsilon_3^2+\epsilon_4^2 +\epsilon_5^2+\epsilon_6^2+\epsilon_7^2)/7$$

How best line and residuals change with a? (b=0 for simplicity)



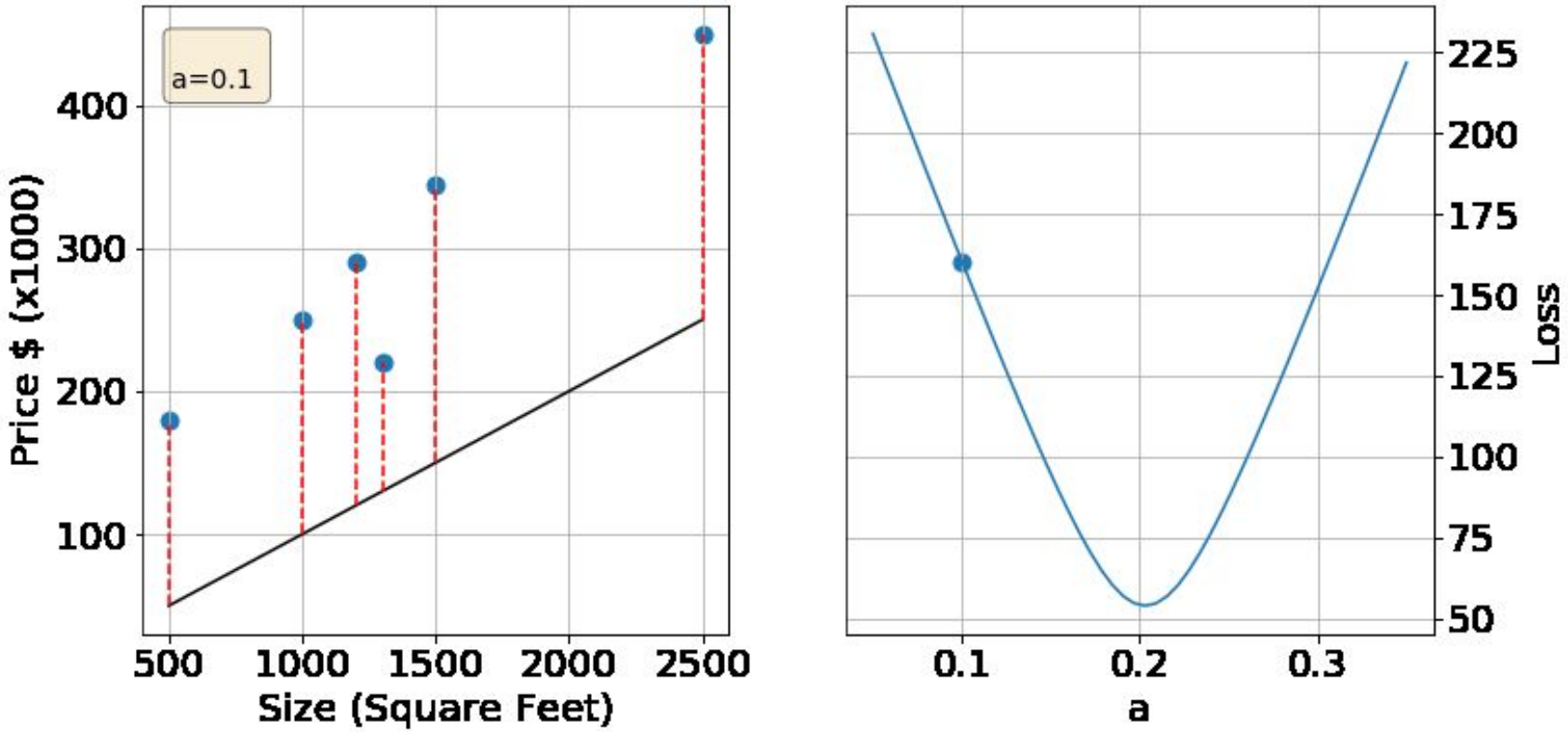
Loss Function



How to spot the best a value?

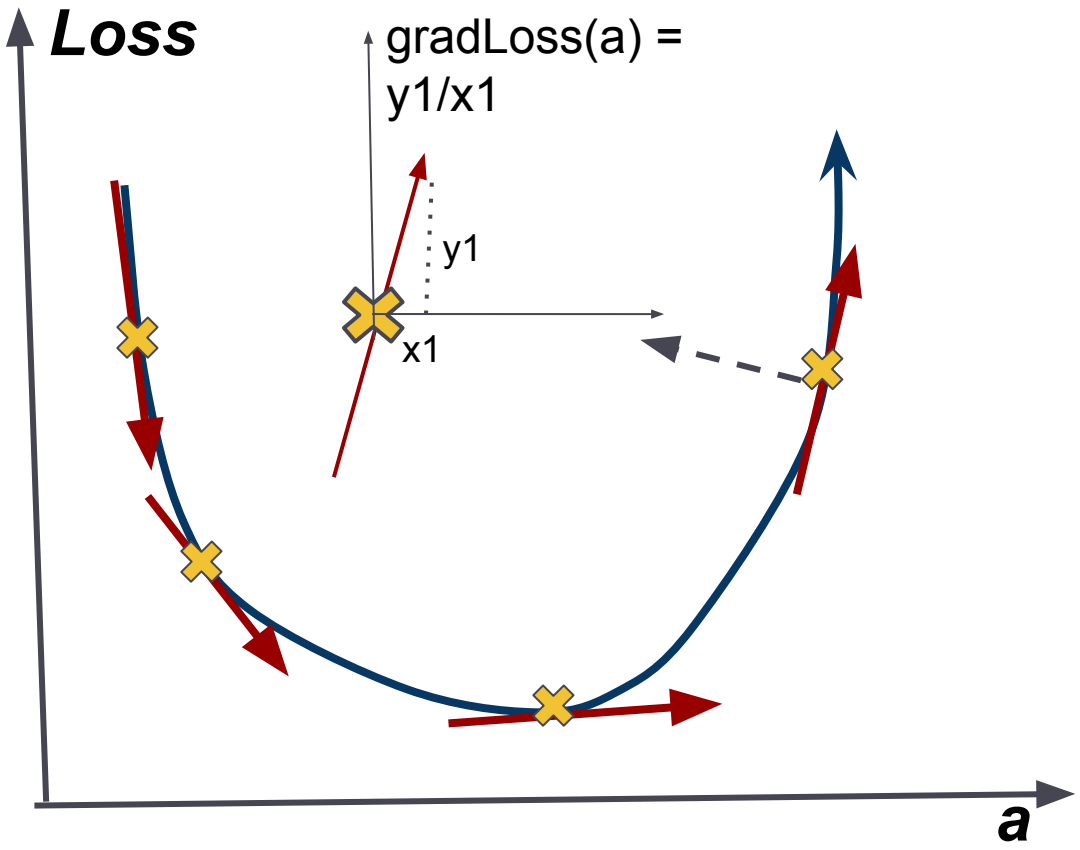
Grid Search

Try a set of designated a values.
Choose the one returning the minimum loss.



a = [0.1, 0.15, 0.2, 0.25, 0.3]

Gradient Descent



1. Randomly initialize a

   $a = a_0$

2. Update a

   $a := a - \Delta \text{gradLoss}(a)$

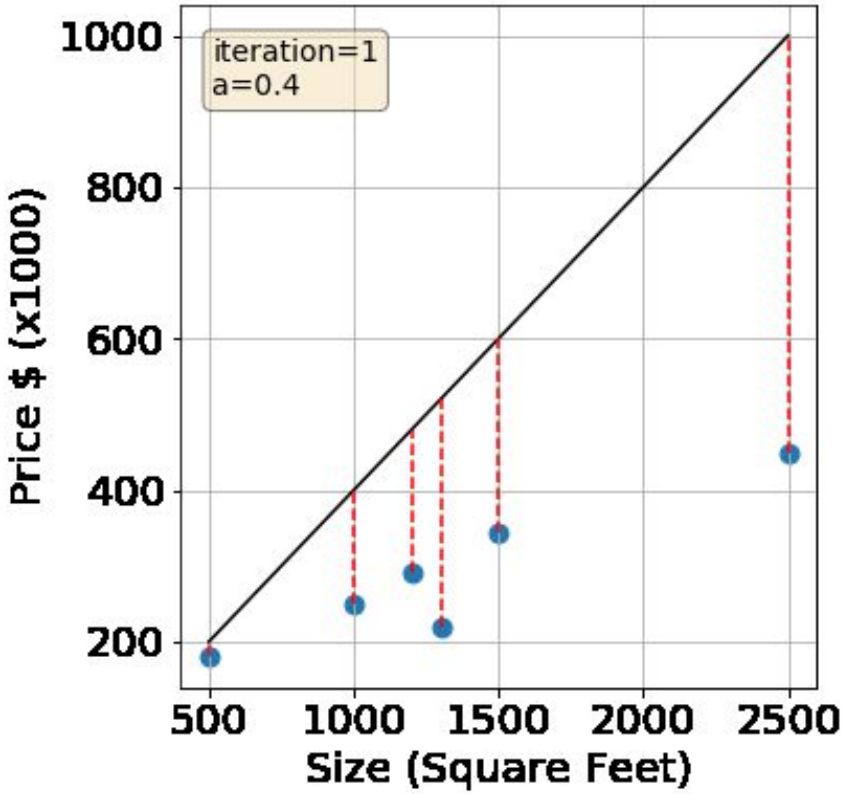
   Δ: step size

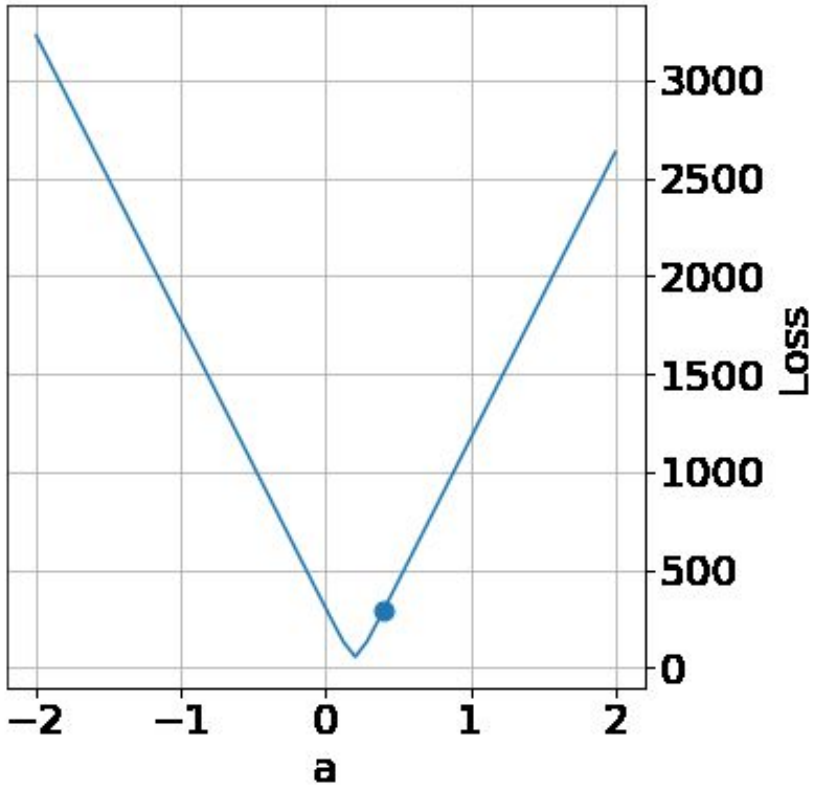
Iterate step 2

Gradient descent in action

a_0=0.4 Δ=1x10^{-6}

a_{0opt}=? Δ_{opt}=?

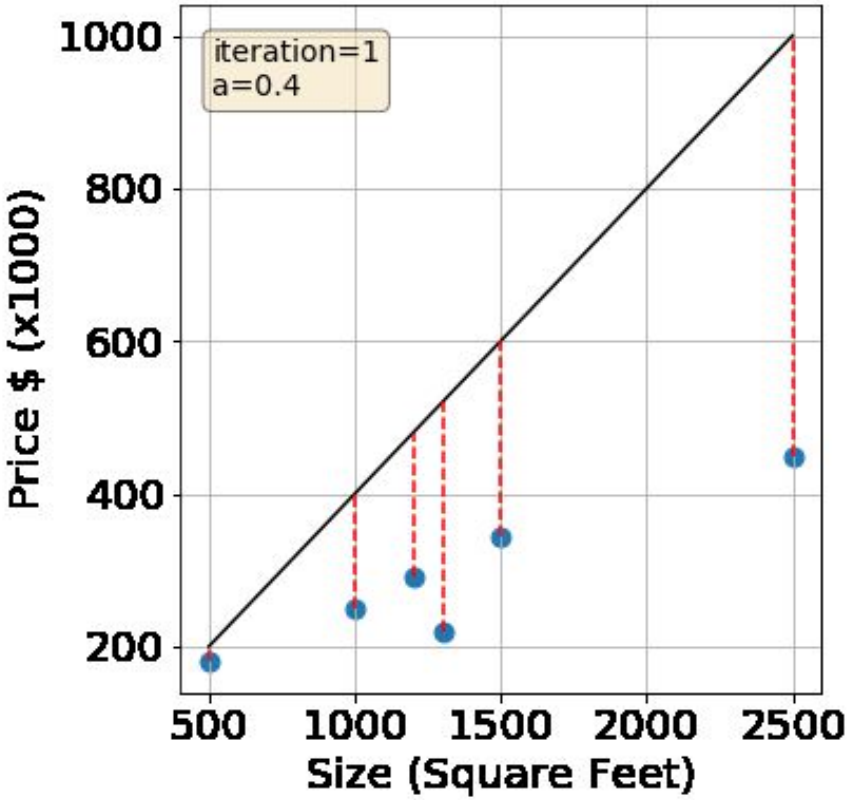


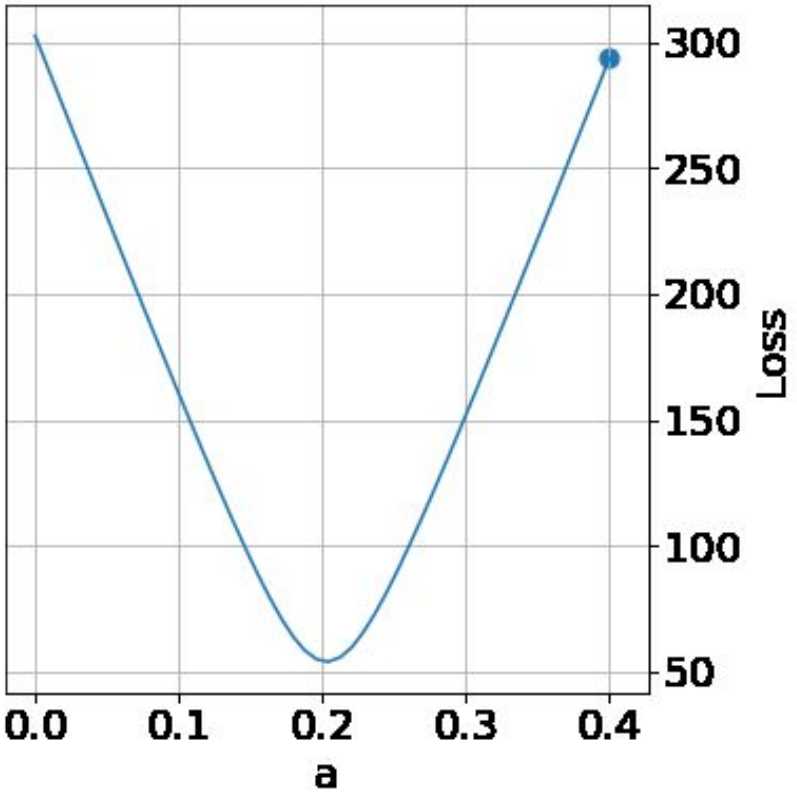


Gradient descent: varying Δ

a_0=0.4 Δ=1x10^{-7}

Takes 17 iterations to complete.

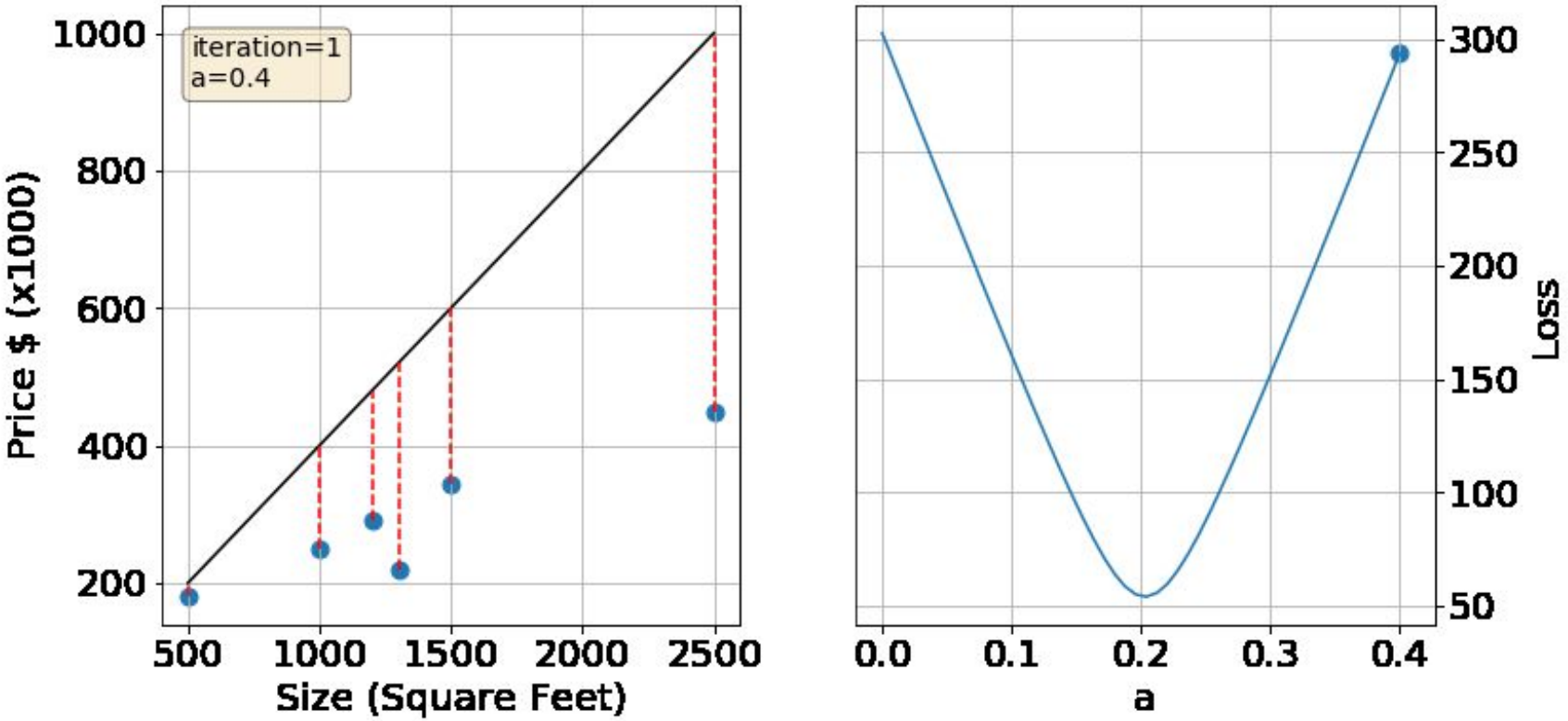




Gradient Descent: varying Δ

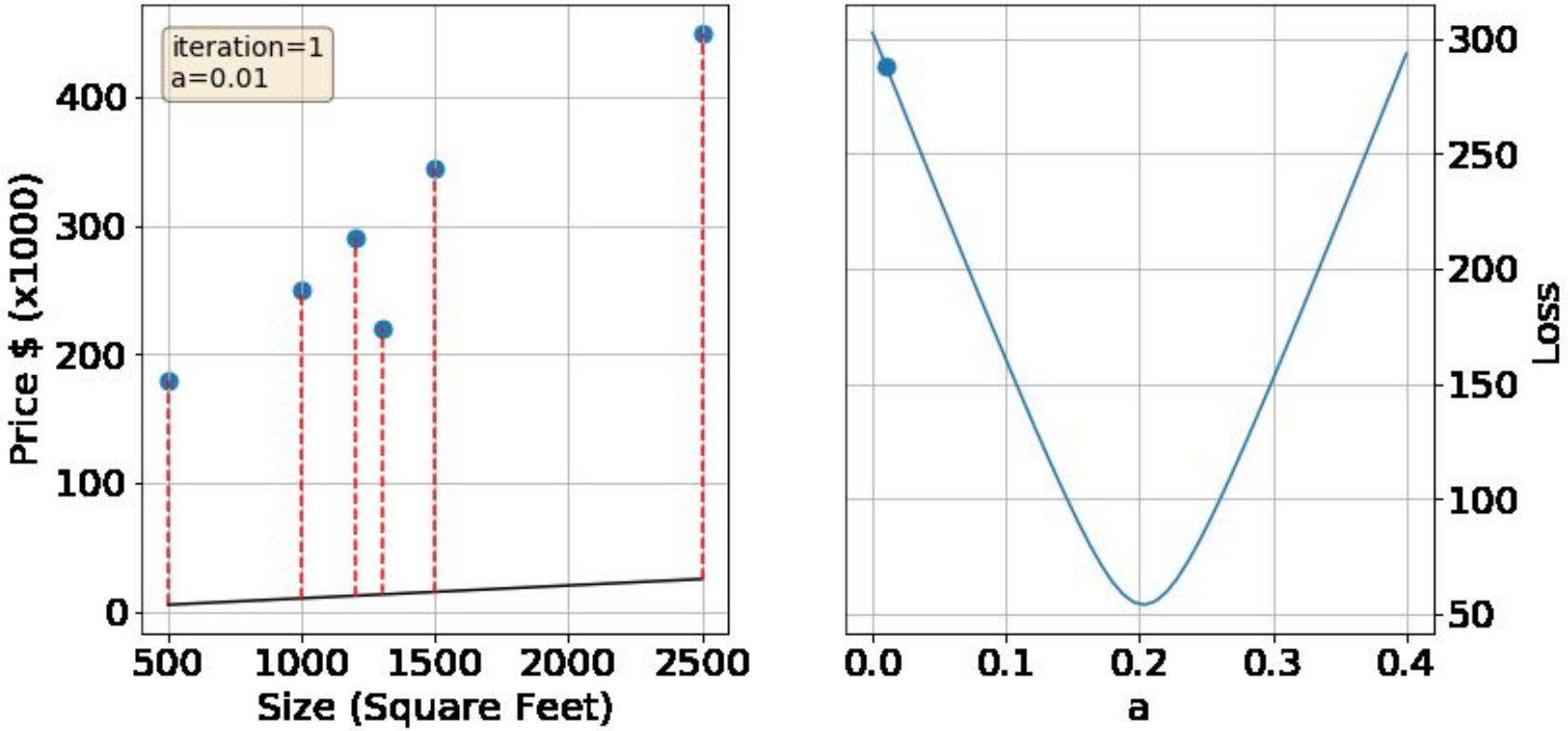
a_0=0.4 $\Delta=3x10^{-7}$ **Takes 6 iterations to complete.**

Important for large datasets!



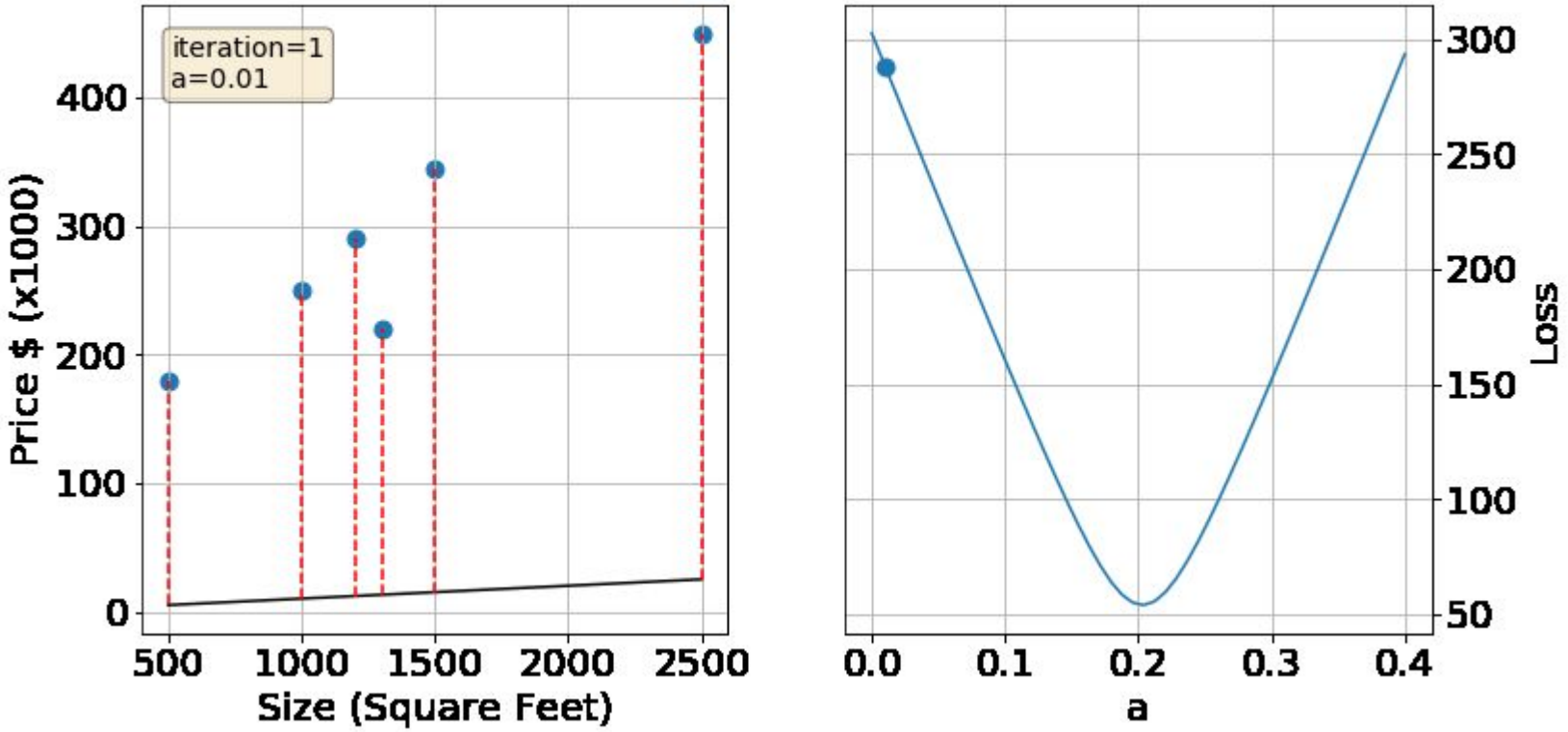
Gradient Descent: varying a_o

a_0=0.01 Δ=3x10^{-7}

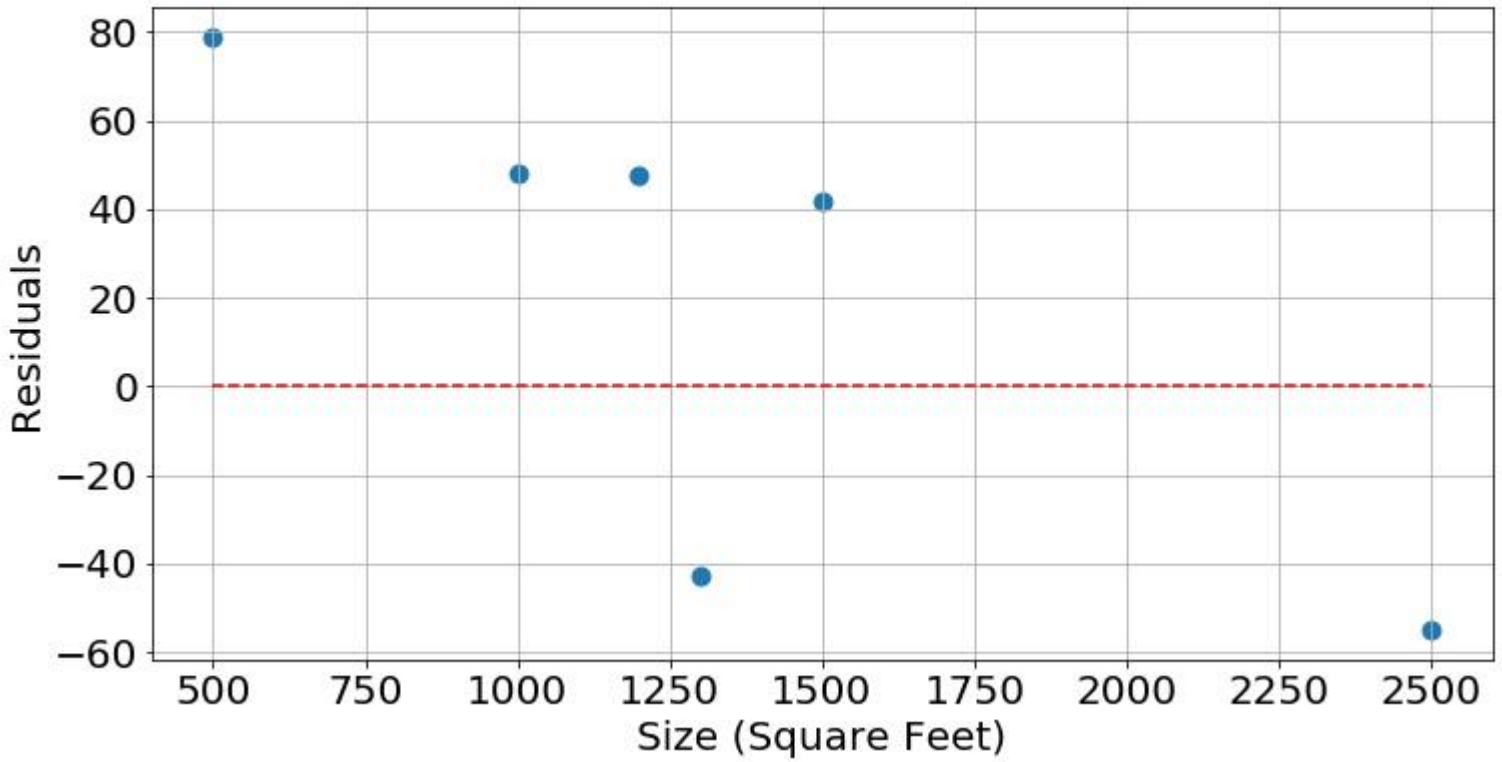


More Efficient Way : Gradient descent

a_0=0.01 Δ=3x10^{-7} , tol = 1e^{-4}

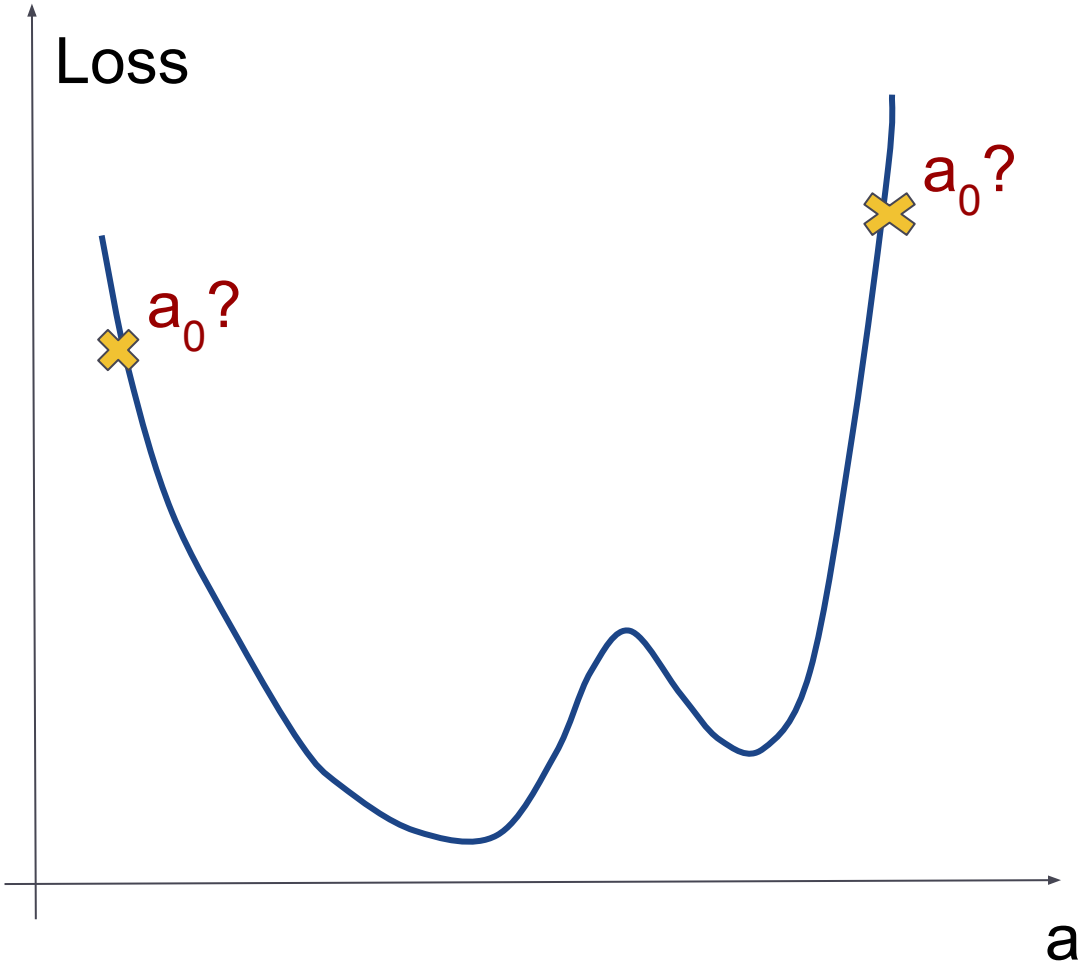


How residuals look for best line?



What about local minima?

What is the optimal solution to the loss function?



Assume a sufficiently small step size to avoid overshoot.

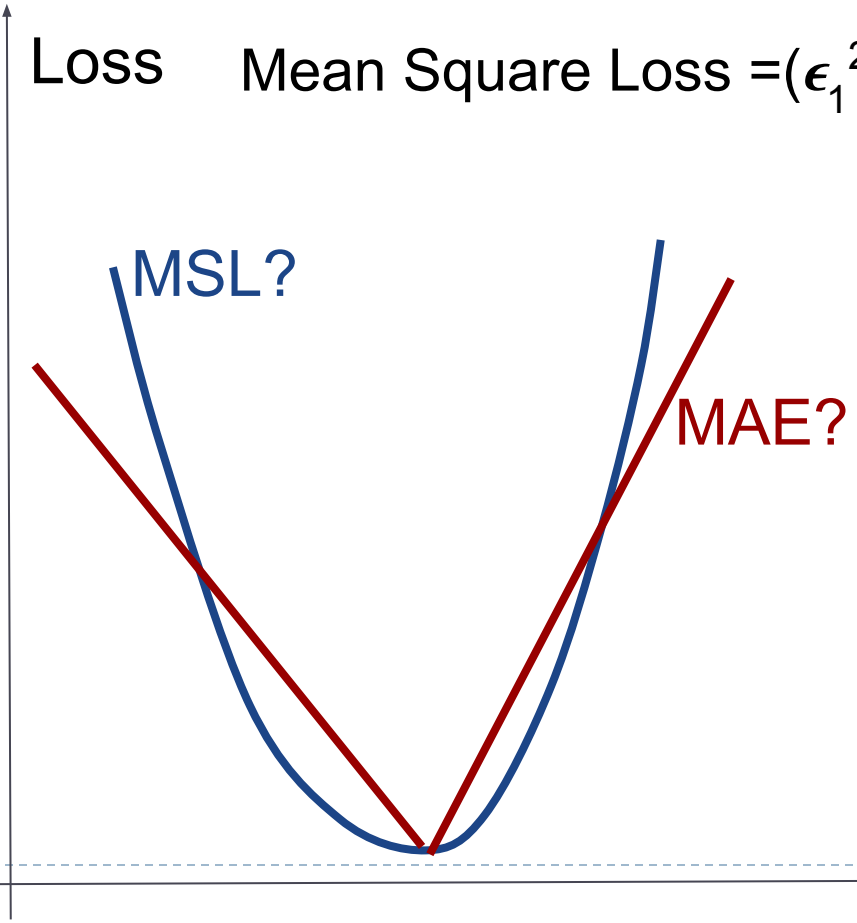
Answer:
Multiple
Initialization

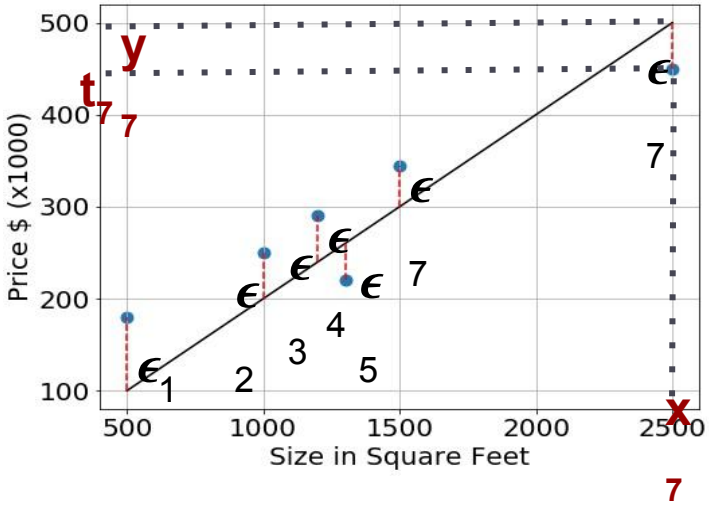
Choosing a Loss Function

Mean Error = $(\epsilon_1 + \epsilon_2 + \epsilon_3 + \epsilon_4 + \epsilon_5 + \epsilon_6 + \epsilon_7)/7$

Mean Absolute Error = $(|\epsilon_1| + |\epsilon_2| + |\epsilon_3| + |\epsilon_4| + |\epsilon_5| + |\epsilon_6| + |\epsilon_7|)/7$

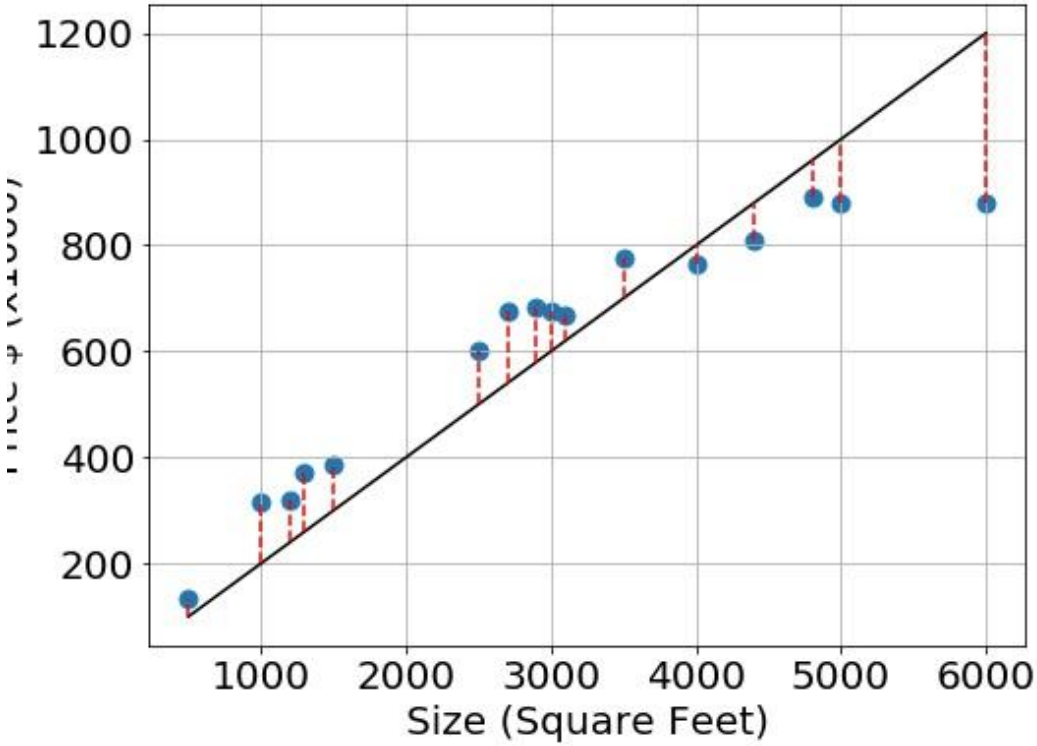
Mean Square Loss = $(\epsilon_1^2 + \epsilon_2^2 + \epsilon_3^2 + \epsilon_4^2 + \epsilon_5^2 + \epsilon_6^2 + \epsilon_7^2)/7$

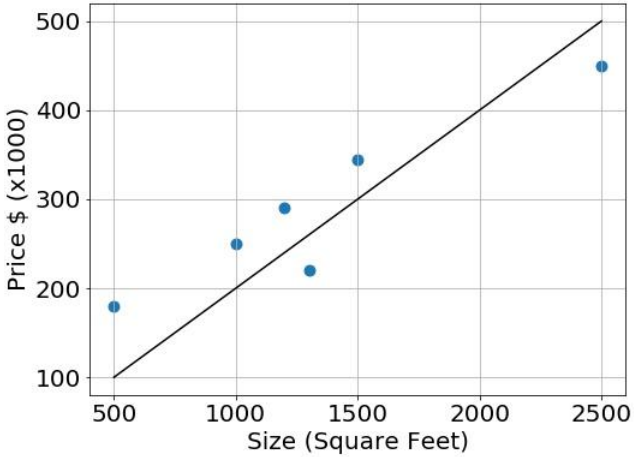




Larger houses added to the dataset

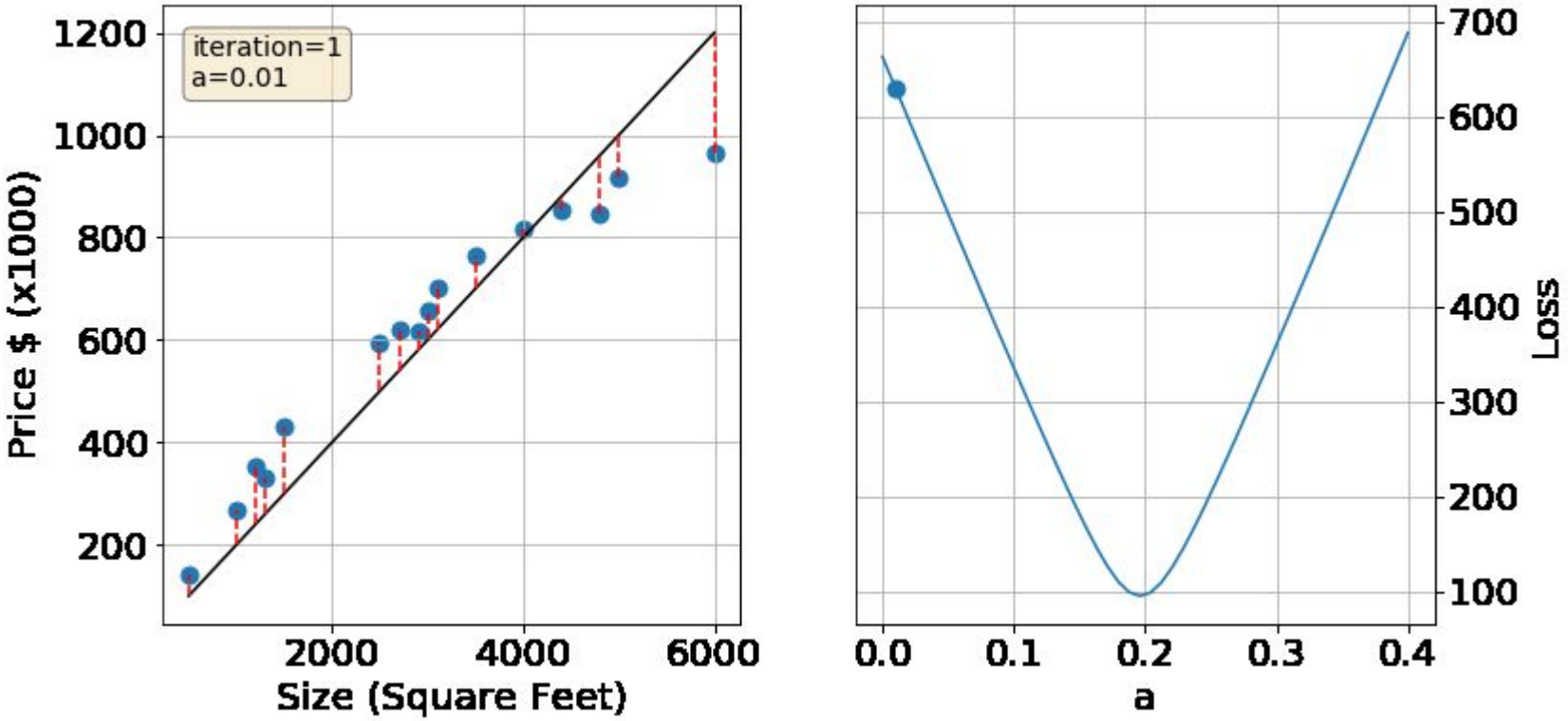
How is the new dataset look different than before ?





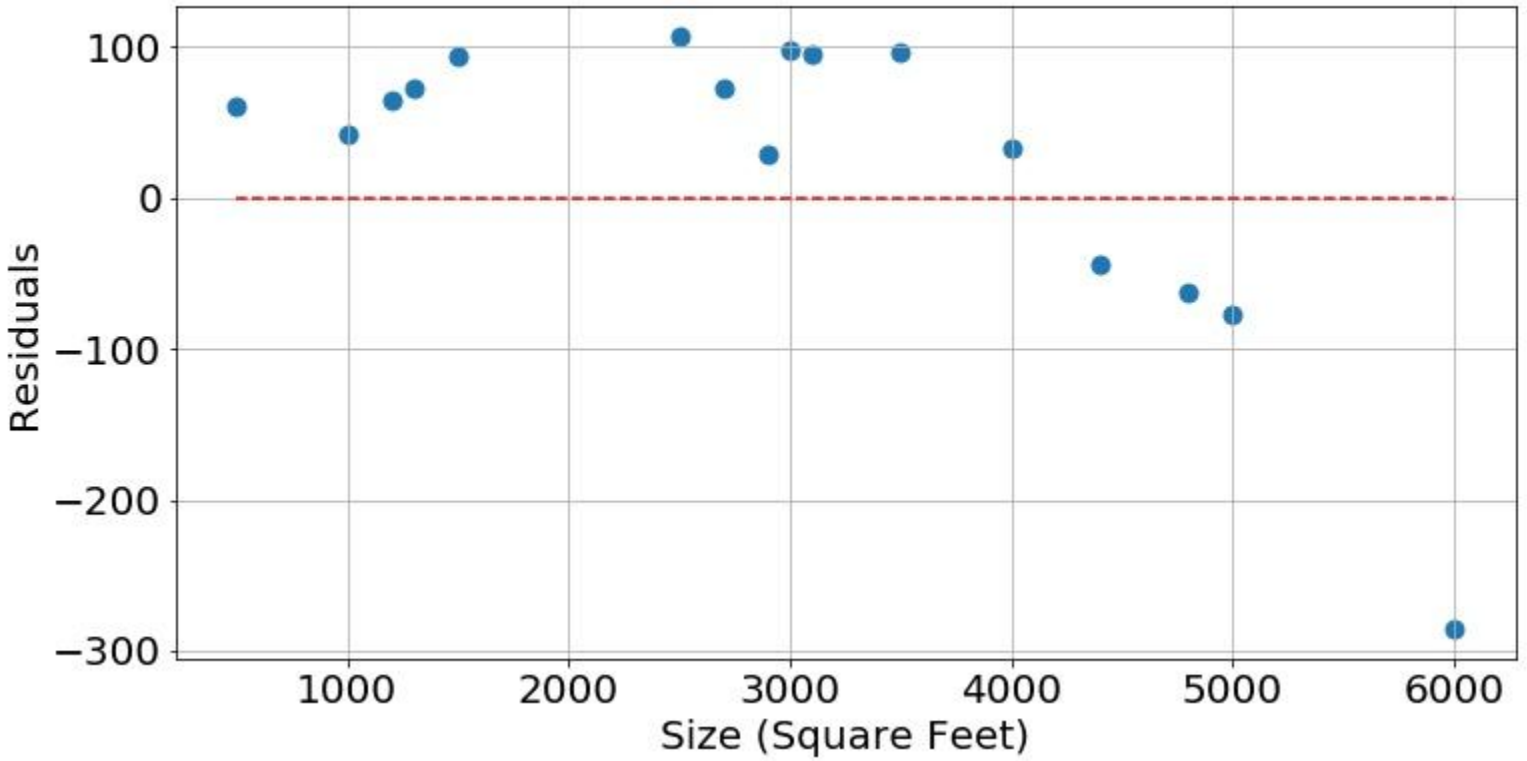
Model with Linear Regression

$$y = ax+b$$



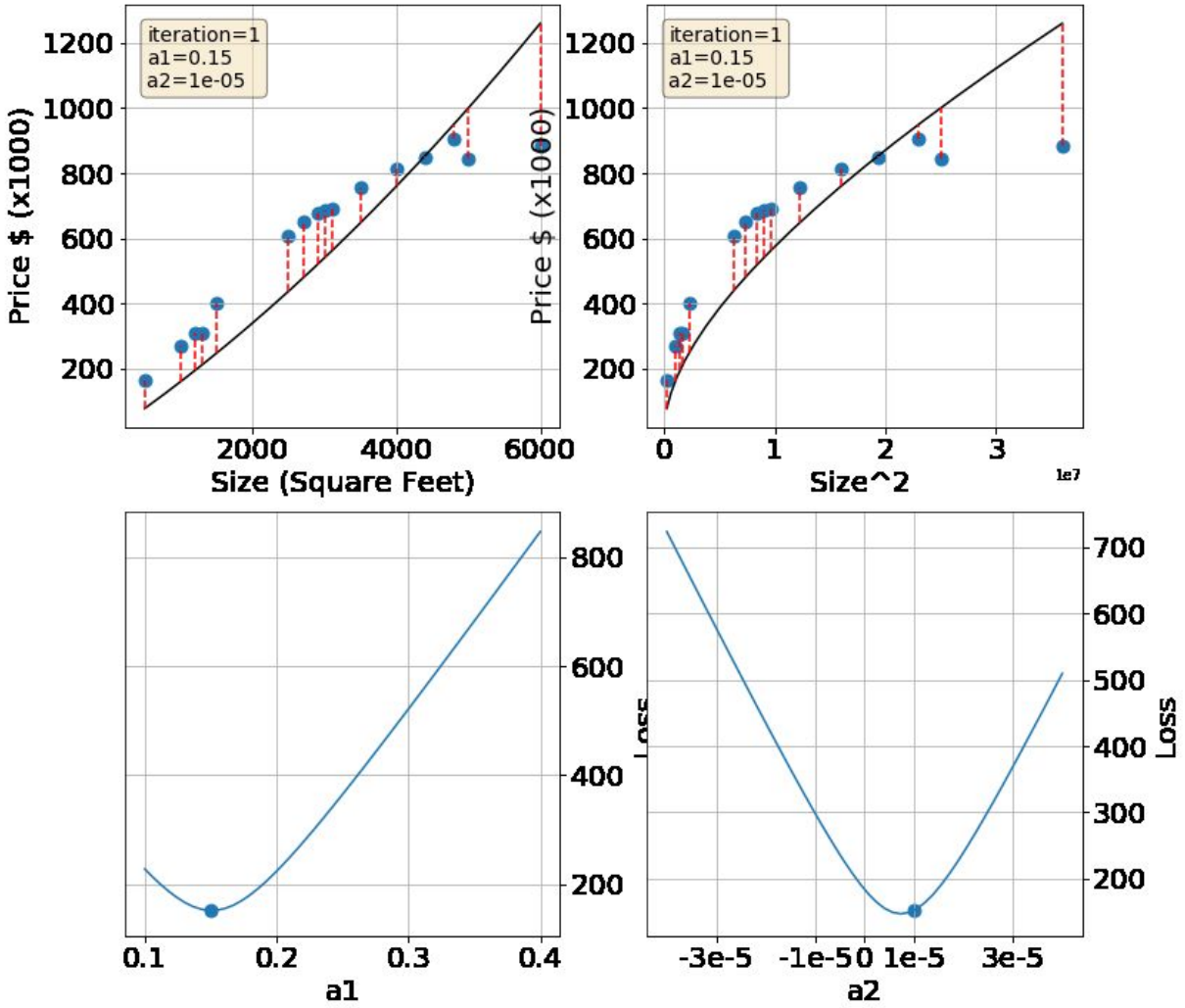
Residuals I

Gets larger for larger houses!



Our assumption of linearity fails!

Polynomial Linear Regression



Polynomial regression is a special case of multiple linear regression (multiple model inputs)

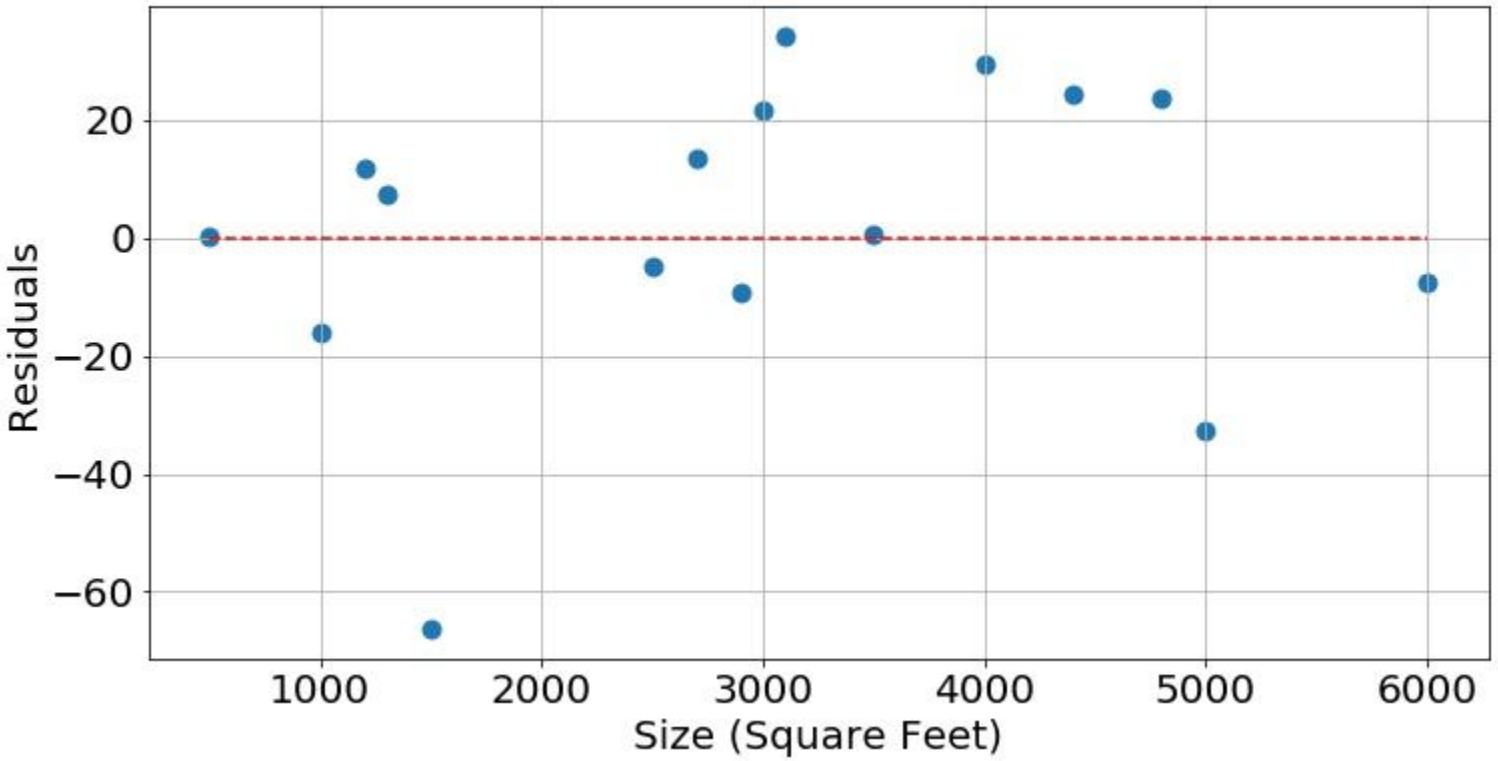
$$y = a_1x_1+a_2x_2+b$$

$$x_1=x\ x_2=x^2$$

$$y = a_1x+a_2x^2+b$$

Residuals II

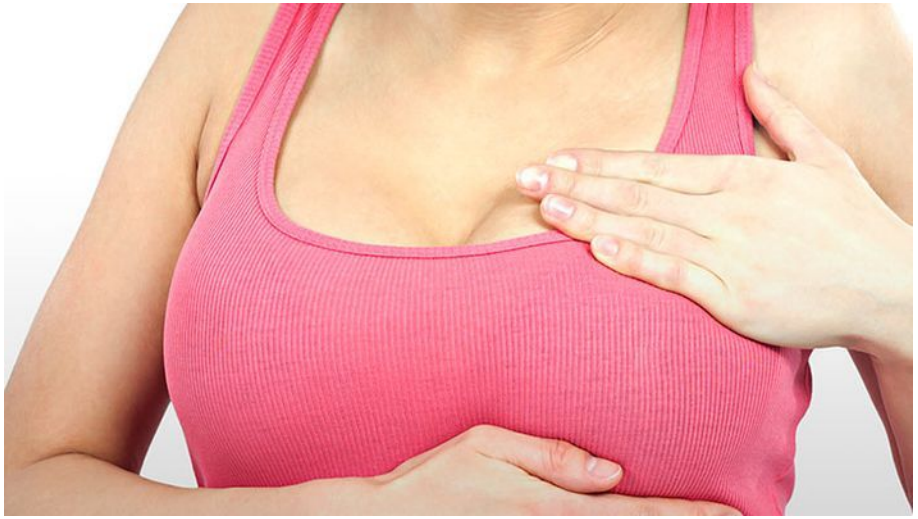
Residuals nicely distributed around 0.



Lecture 3: Machine Learning Techniques II
11:10-12:00

Logistic Regression

Predicting Breast Cancer



Pathology Results

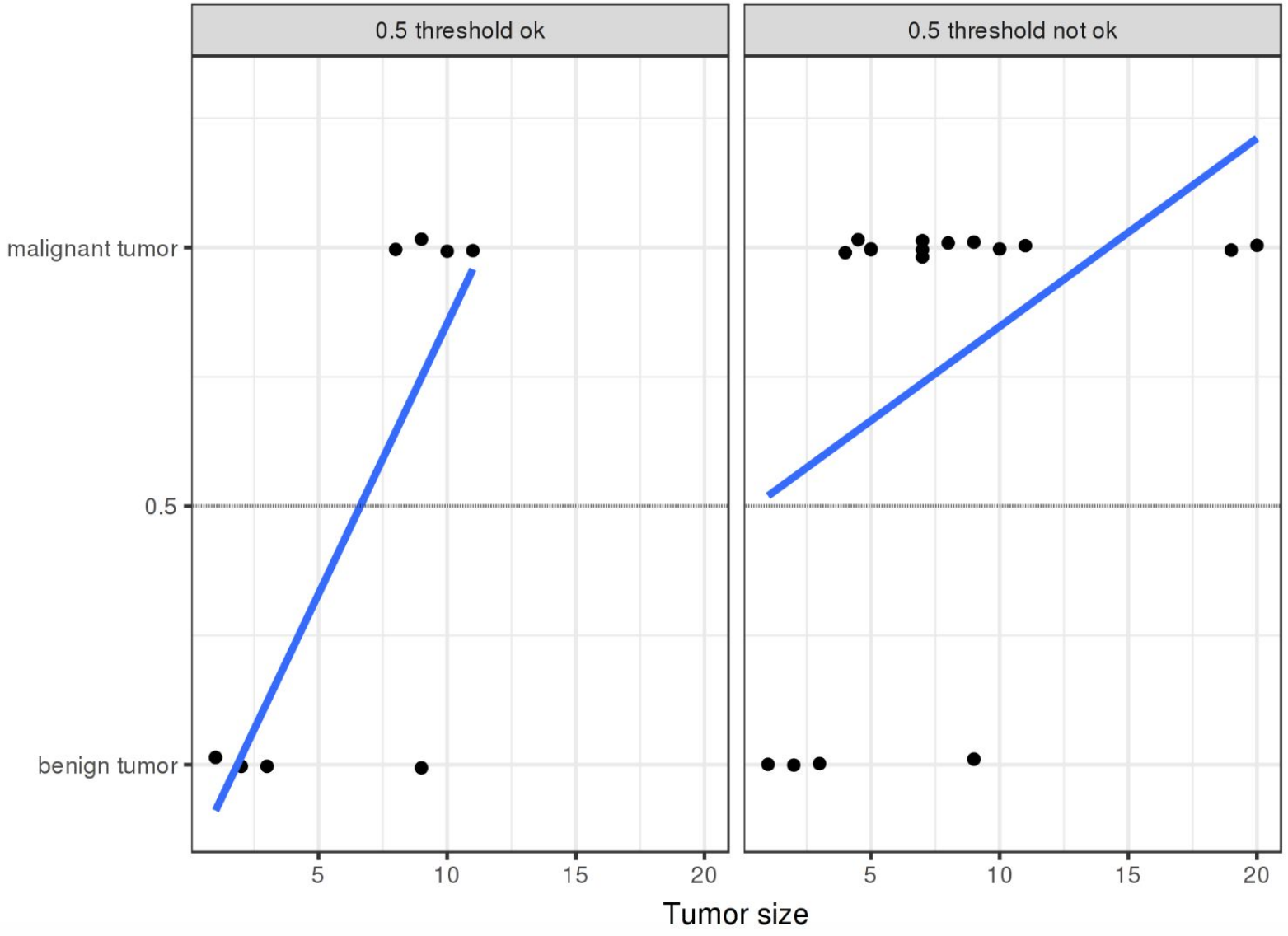
Tumor Size	Cancer
1	No
2	No
3	No
8	Yes
9	Yes
11	Yes
12	Yes

New Patients

Tumor Size	Cancer
2.5	?
5.5	?

Predicting breast cancer is a classification problem.
One ML algorithm to solve this problem is Logistic Regression.

What is wrong with LR for classification?



Probabilistic Approach

$$P(t=1)/P(t=0) = P(t=1)/1-P(t=1) = e^{ax+b}$$

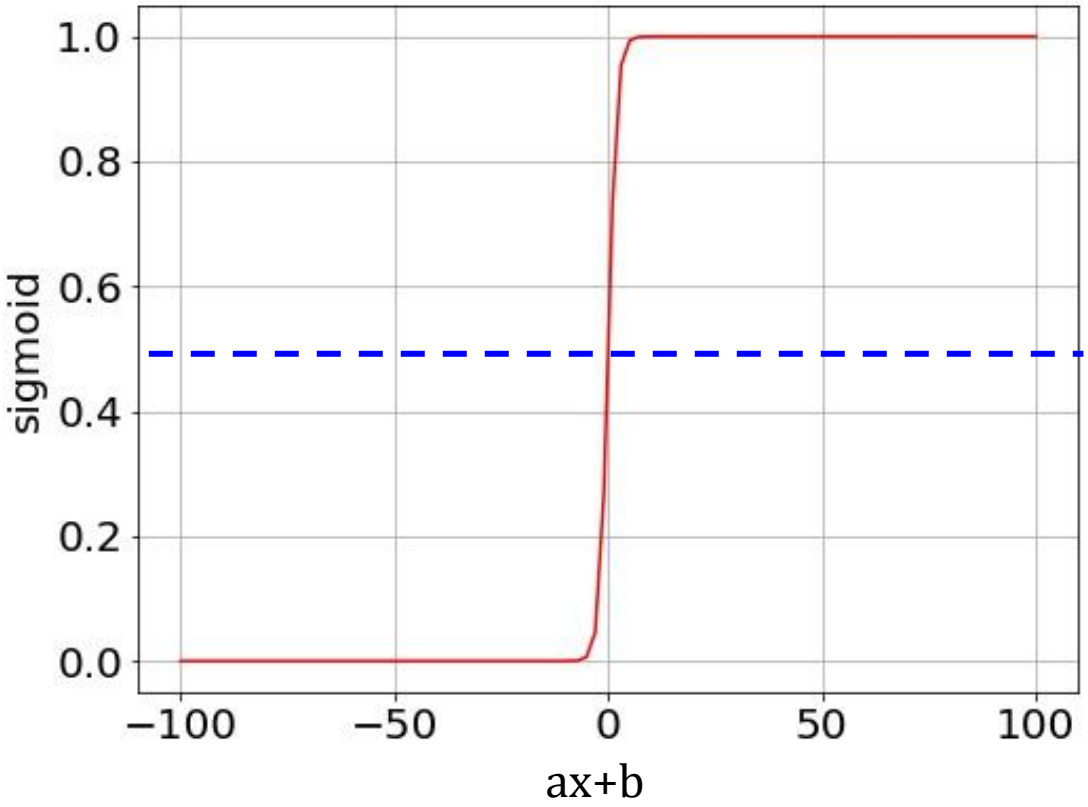
Mapping ax+b to probability of observing positive class

$$P(t=1) = 1/(1+e^{-(ax+b)})$$

Also called sigmoid

More than one feature? Ex. tumor size and patient age

Sigmoid

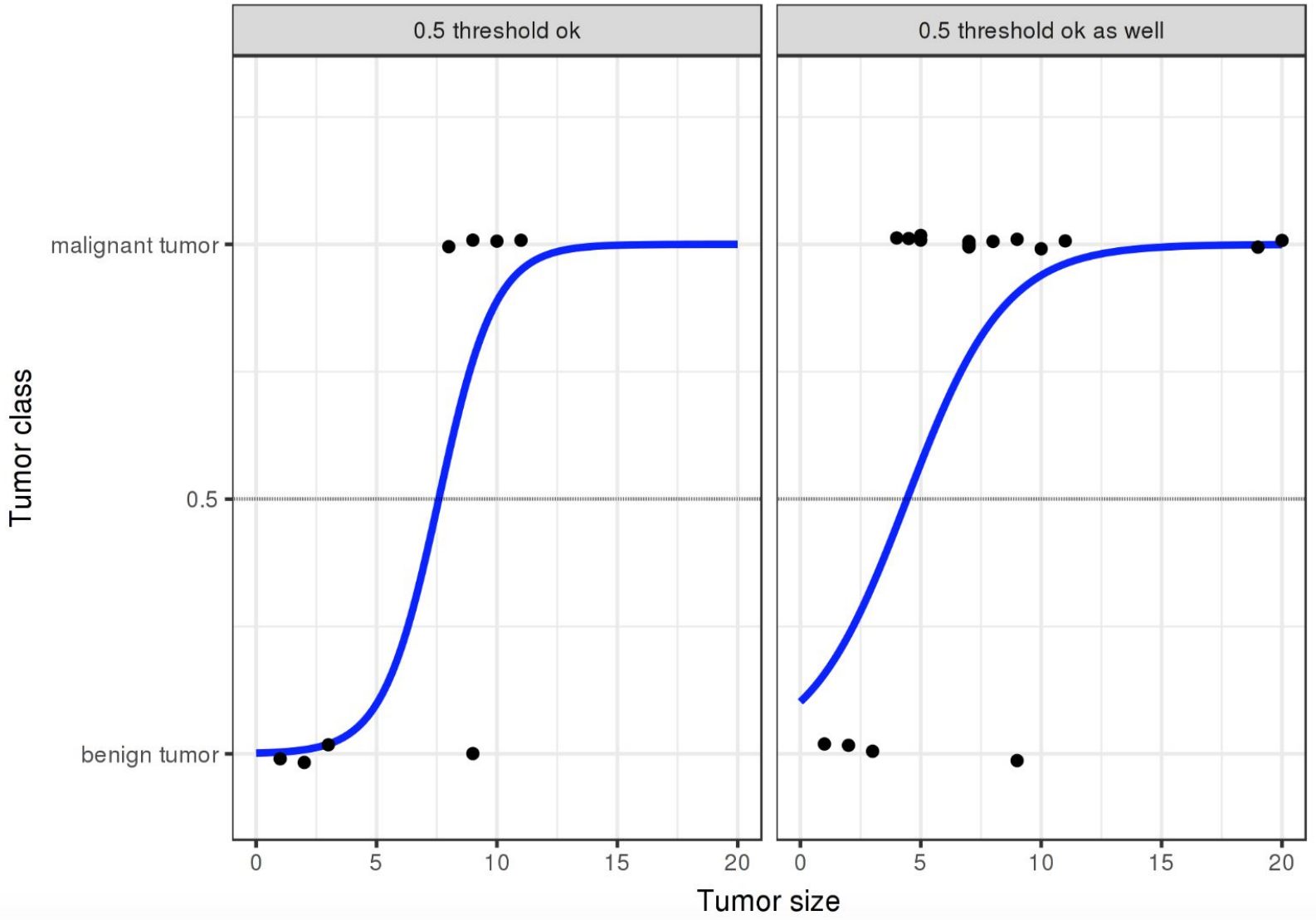


Decision boundary
sigmoid = 0.5

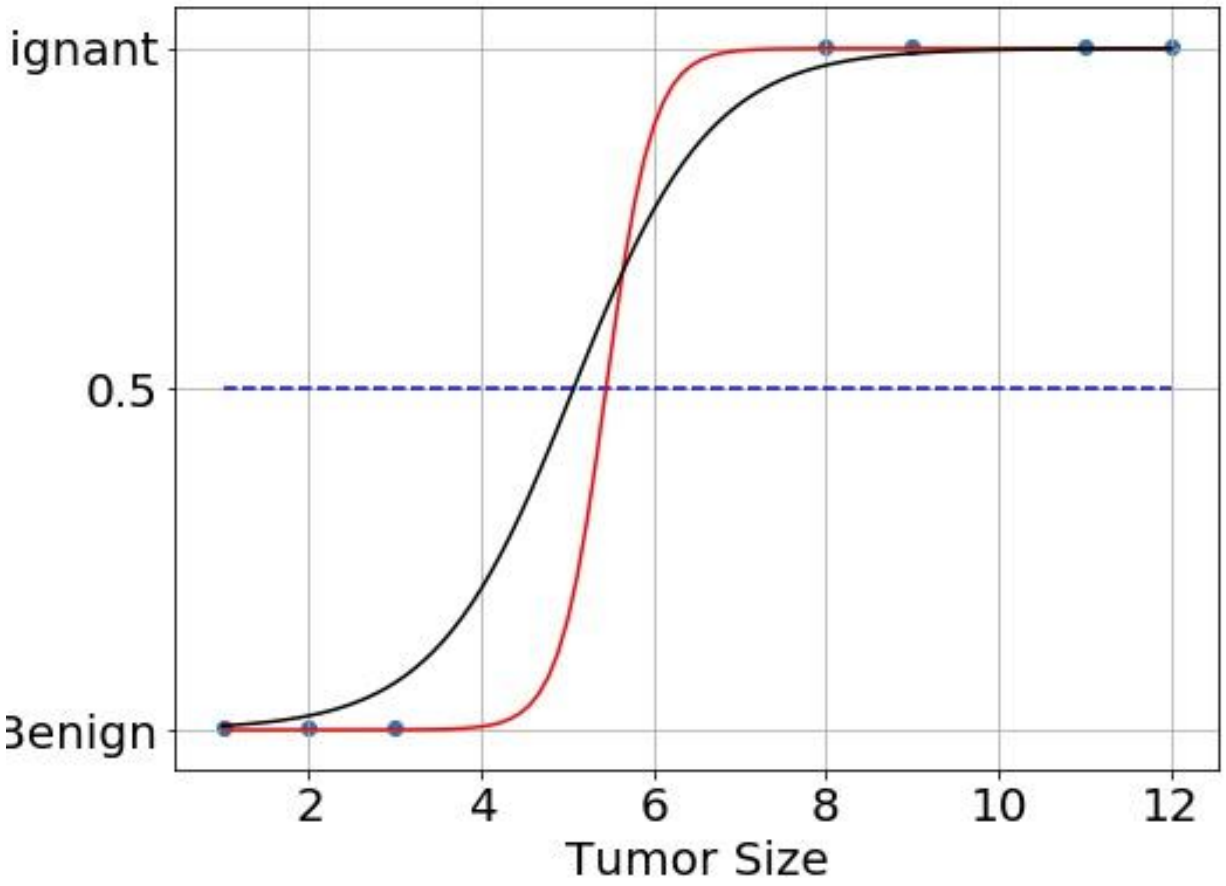
y>0 => 0.5<sigmoid<1

y<0 => 0<sigmoid<0.5

Sigmoid

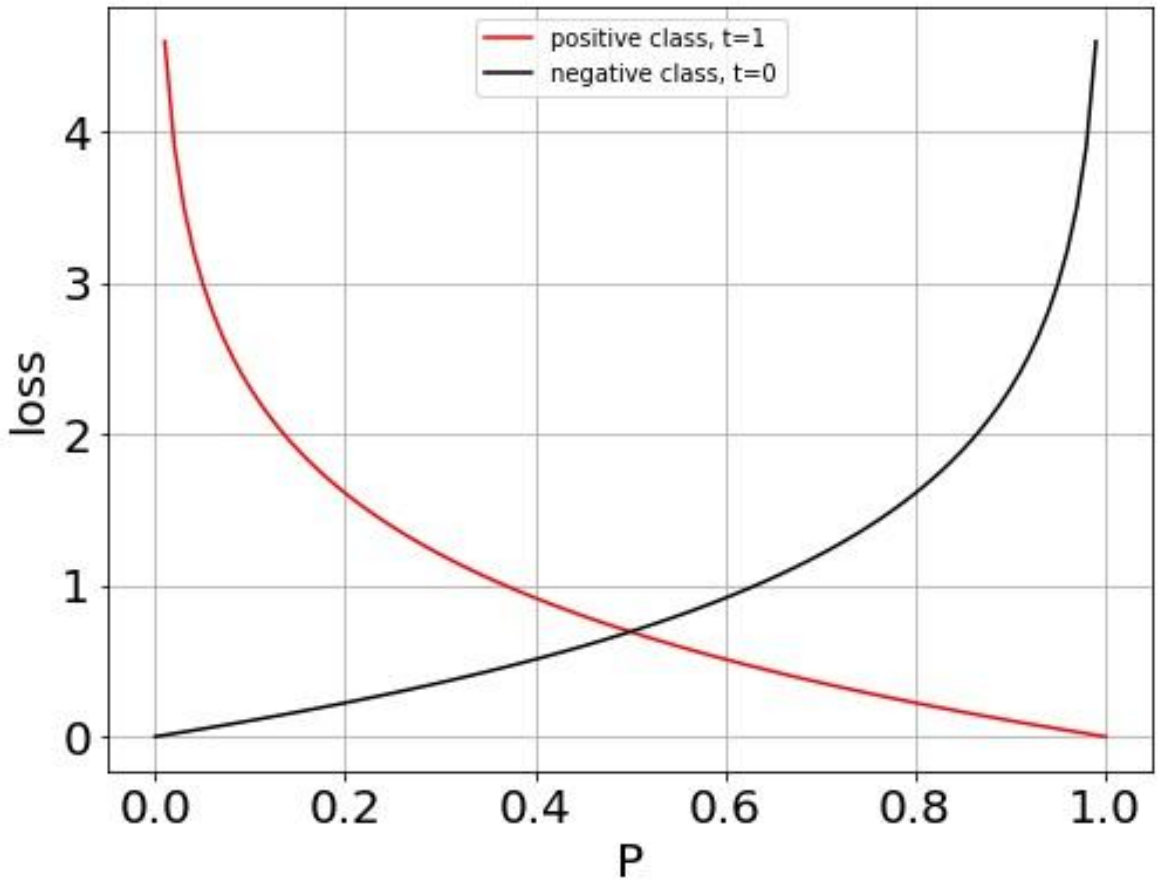

0.5 threshold ok
0.5 threshold ok as well
malignant tumor
0.5
benign tumor
Tumor class
0
5
10
15
20
Tumor size

Best Model?



Log Loss (Cross Entropy)

If t=1 [positive class] , ***loss = -log p***
If t=0 [negative class], ***loss =-log(1-p)***



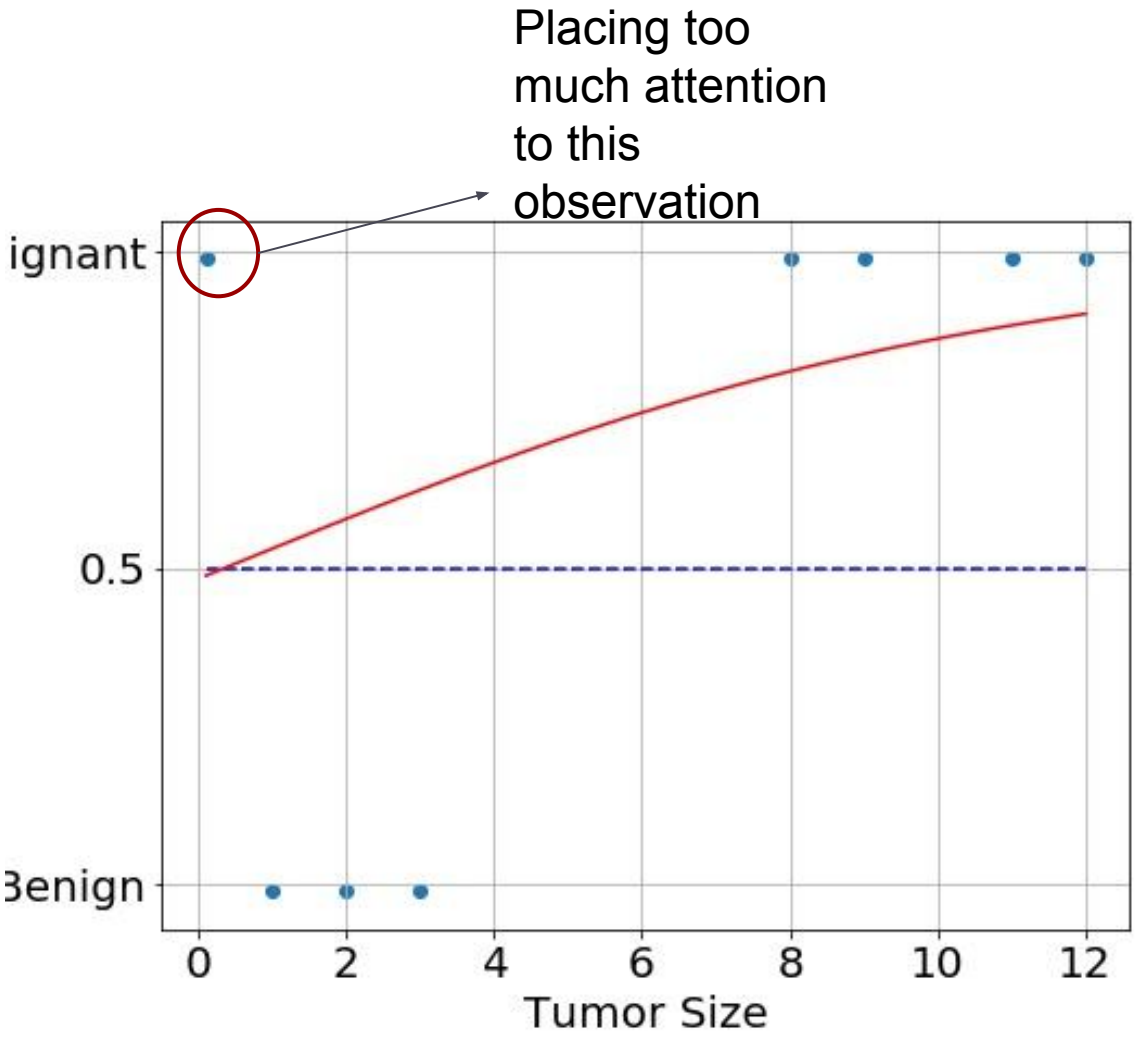
Log Loss (Cross Entropy)

loss = -tlog(p)+(1-t)log(1-p) [combined]
t: actual value [0 or 1]

Minimization of log loss means choosing the model parameters (a,b) which maximize the probability of observation of training data.

* Derivation requires probability theory and beyond our scope.
** I would highly recommend interested attendees to read about Bayesian approach to logistic regression.

What if data not perfectly separable?



How to penalize tuning to peculiarities?

Regularization

Regularization parameter

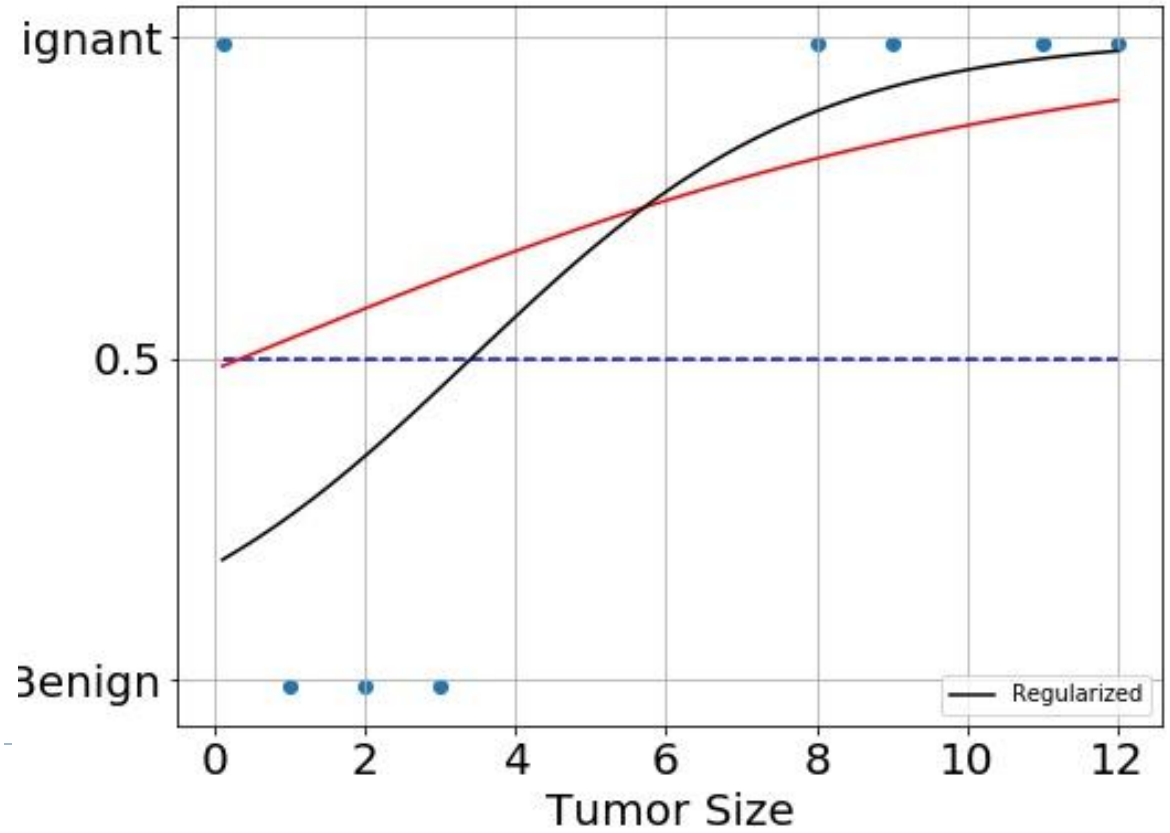
$$loss = -t\log(p)+(1-t)\log(1-p)+Ca^2$$

- Regularization does NOT improve the performance on the data set that the algorithm used to learn the model parameters (feature weights).
- It can improve the generalization performance, i.e., the performance on new, unseen data, which is exactly what we want.

Regularization

- Regularization does NOT improve the performance on the data set that the algorithm used to learn the model parameters (feature weights).

- It can improve the generalization performance, i.e., the performance on new, unseen data, which is exactly what we want.



Is it always fair to set threshold = 0.5?

Task1: Predicting whether a suspect is criminal or not?

You identified a suspect is 55% likely to be criminal.

Would you send this person to jail?

Task2: Predicting breast cancer

You identified a tumor is 40% likely to be malignant.

Would you deem it benign and send the patient home?

Classification Performance (Derived) Metrics

Want to compare t two different classification algorithms, Logisitc Regression vs. Support Vector Machines?

Confusion Matrix

		Predicted: Positive	Predicted: Negative
Actual	Positive	True Positive TP	False Negative FN
	Negative	False Positive FP	True Negative TN

How many items predicted correctly?

Accuracy = (TP+TN)/(TP+FP+TN+FN)

How many of the selected items relevant?

Precision=TP/(TP+FP)

How many of the relevant(positive) items selected ?

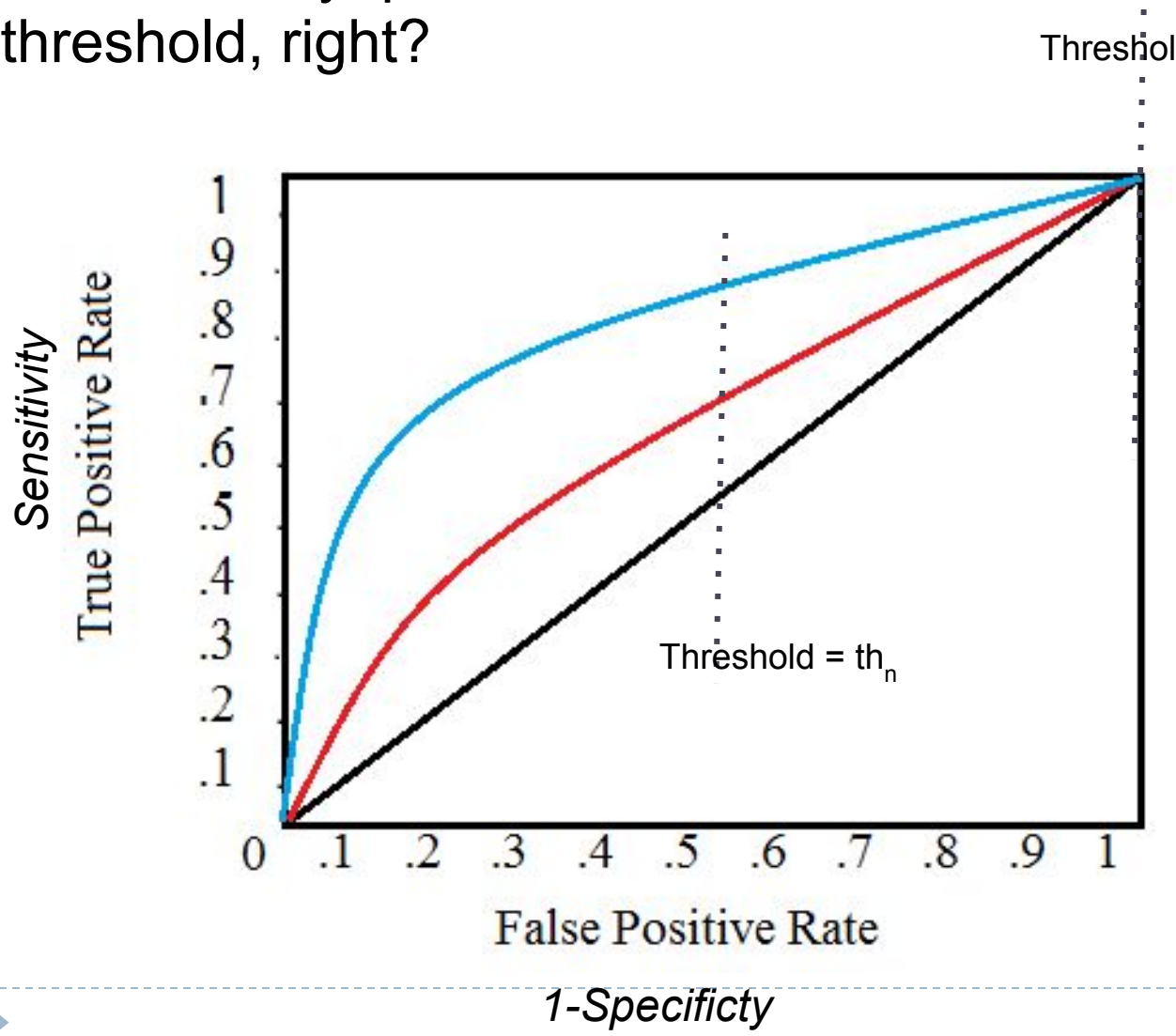
Recall(Sensitivity)= TP/(TP+FN)

How many of the nonrelevant items selected ?

Specificity= TN/(TN+FP)

Receiver Operator Characteristics (ROC) Curve

But accuracy, precision and recall is all based on a threshold, right?



Area under the receiver operator characteristics curve (AUROC)

Lecture 4: Model Selection and Evaluation

1:30-2:30

Model Evaluation

The best predictive algorithm is one that has good ***Generalization Ability***. With that, it will be able to give accurate predictions to new and previously unseen data.

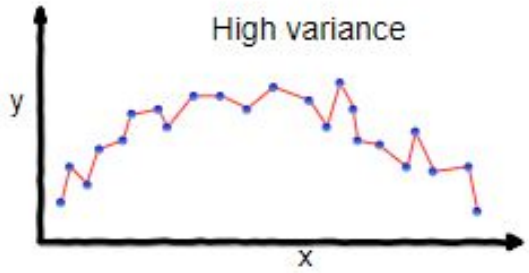
Bias-variance tradeoff is a concept in machine learning which refers to the problem of minimizing two error sources at the same time

Bias variance tradeoff

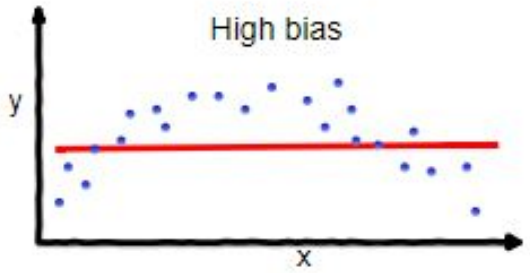
High Bias results from ***Underfitting*** the model (simple model). This usually results from erroneous assumptions, and cause the model to be too general.

High Variance results from ***Overfitting*** the model (complex model), and it will predict the training dataset very accurately, but not with unseen new datasets. This is because it will fit even the slightest noise in the dataset.

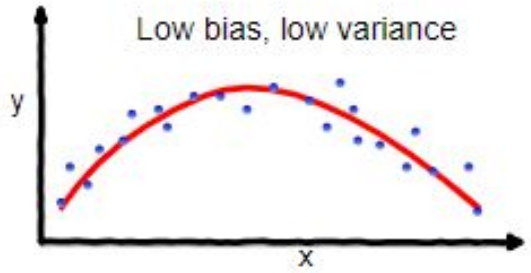
Bias variance tradeoff



overfitting



underfitting



Good balance

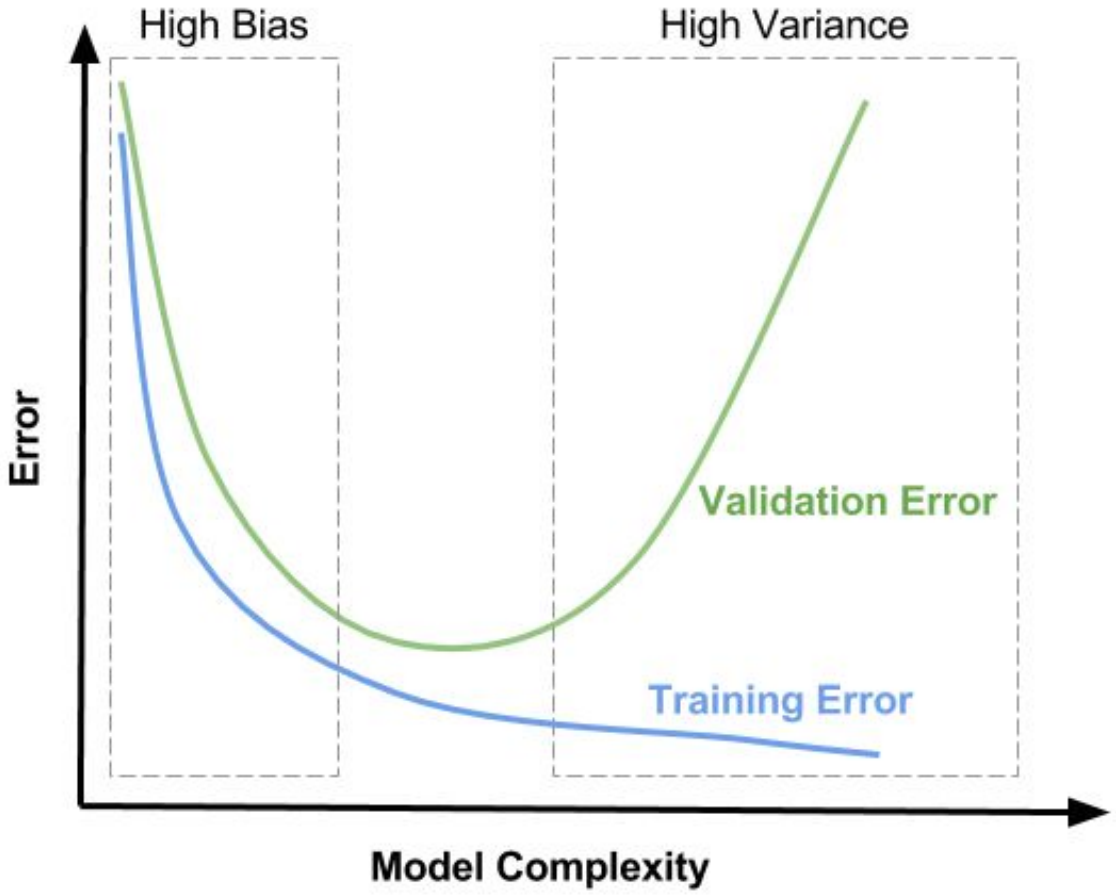
What is the learning error in each case?

Validation and Testing

***Validation Dataset**: The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyperparameters. The evaluation becomes more biased as skill on the validation dataset is incorporated into the model configuration*

***Test Dataset**: The sample of data used to provide an unbiased evaluation of a final model fit on the training dataset.*The Test dataset provides the gold standard used to evaluate the model. It is only used once a model is completely trained(using the train and validation sets). The test set is generally what is used to evaluate competing models

Optimal Model Complexity



Dev/Test Data Distributions

Task: predicting house prices

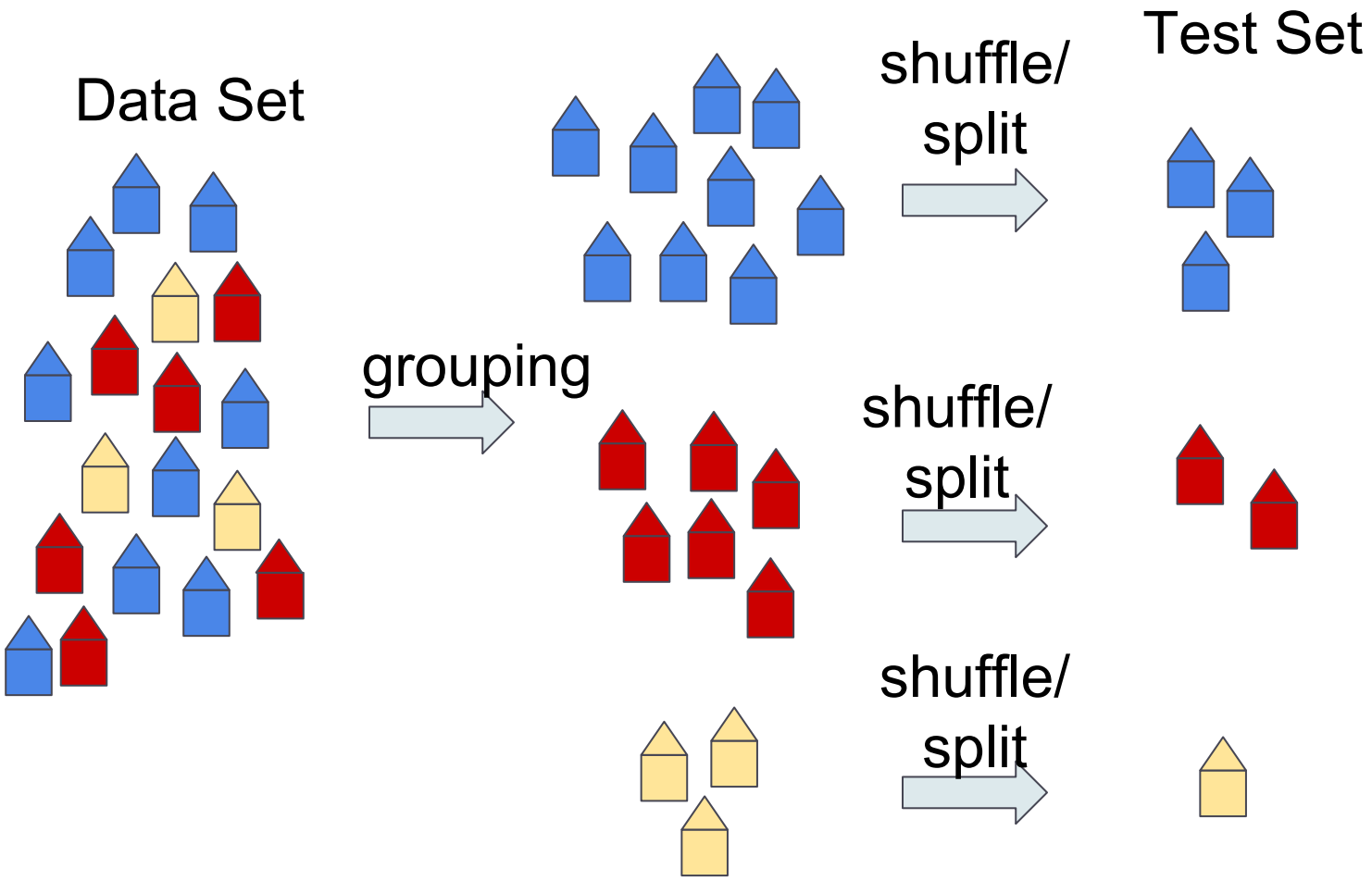
Dev Set

- Portland, OR
- Seattle, WA
- Washington DC

Test Set

- Ann Arbor, MI
- Durham, NC
- Champaign, IL

Stratified Splitting



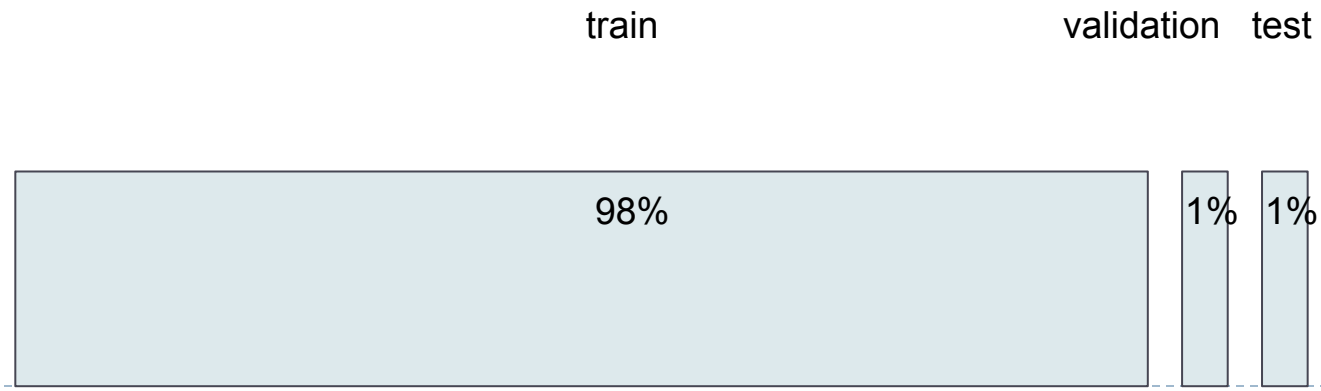
Guideline

Choose a dev set and test set to reflect data you expect to get in the future and consider important to do well on.

Size of train, validation and test sets

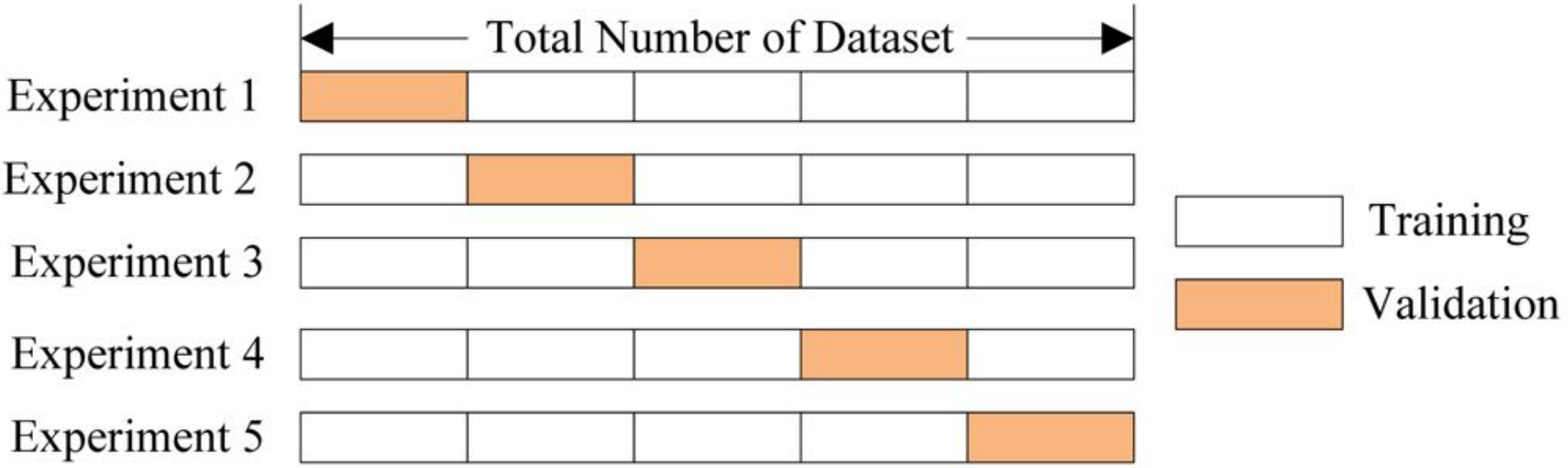
Rule of thumb

Big data > 1M



Cross- validation

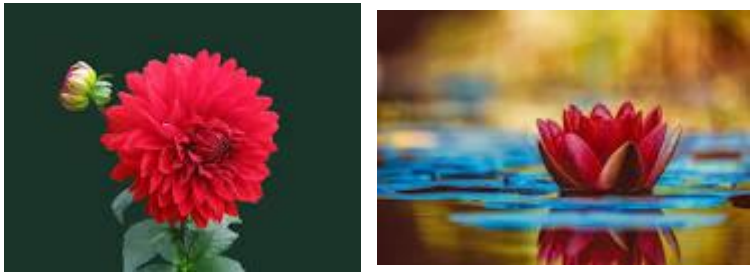
When you have little data ..



Training and testing on Different Distributions

Task: Mobile flower categorization app

Images from web



n=100000

Images from mobile app

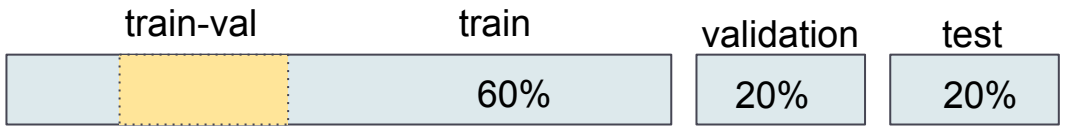
n=10000

How to form train,

validation and test sets?

Bias and Variance with Mismatched Distributions

Training-validation set: same distribution as training set but not used for training



Error	MODEL1	MODEL2	MODEL3
Training Set	1%	10%	1%
Training-Validation Set	9%	11%	1.5%
Validation Set	%10	12%	10%