Web Activity Logs

- Web activity logs are an absurdly rich source of information
- People have examined logs to learn all sorts of things
  - Ideas of what you could learn about someone from their web activity?

Web Activity Logs

- Web activity logs are an absurdly rich source of information
- People have examined logs to learn all sorts of things
  - System failures
  - Security intrusions
  - Buying habits
  - Spelling habits
  - Web browsing behavior
  - Impending disease

POLICY & LAW | US & WORLD | NET NEUTRALITY

Congress just cleared the way for internet providers to sell your web browsing history

Resolution is now off to the president's desk

by Jacob Kastrenakes | Mar 28, 2017, 5:57pm EDT

By THE ASSOCIATED PRESS MARCH 23, 2017, 6:54 P.M. E.D.T.

NEW YORK — The Senate voted to kill Obama-era online privacy regulations , a first step toward allowing internet providers such as Comcast, AT&T and Verizon to sell your browsing habits and other personal information as they expand their own online ad businesses.

Machine Learning Introduction

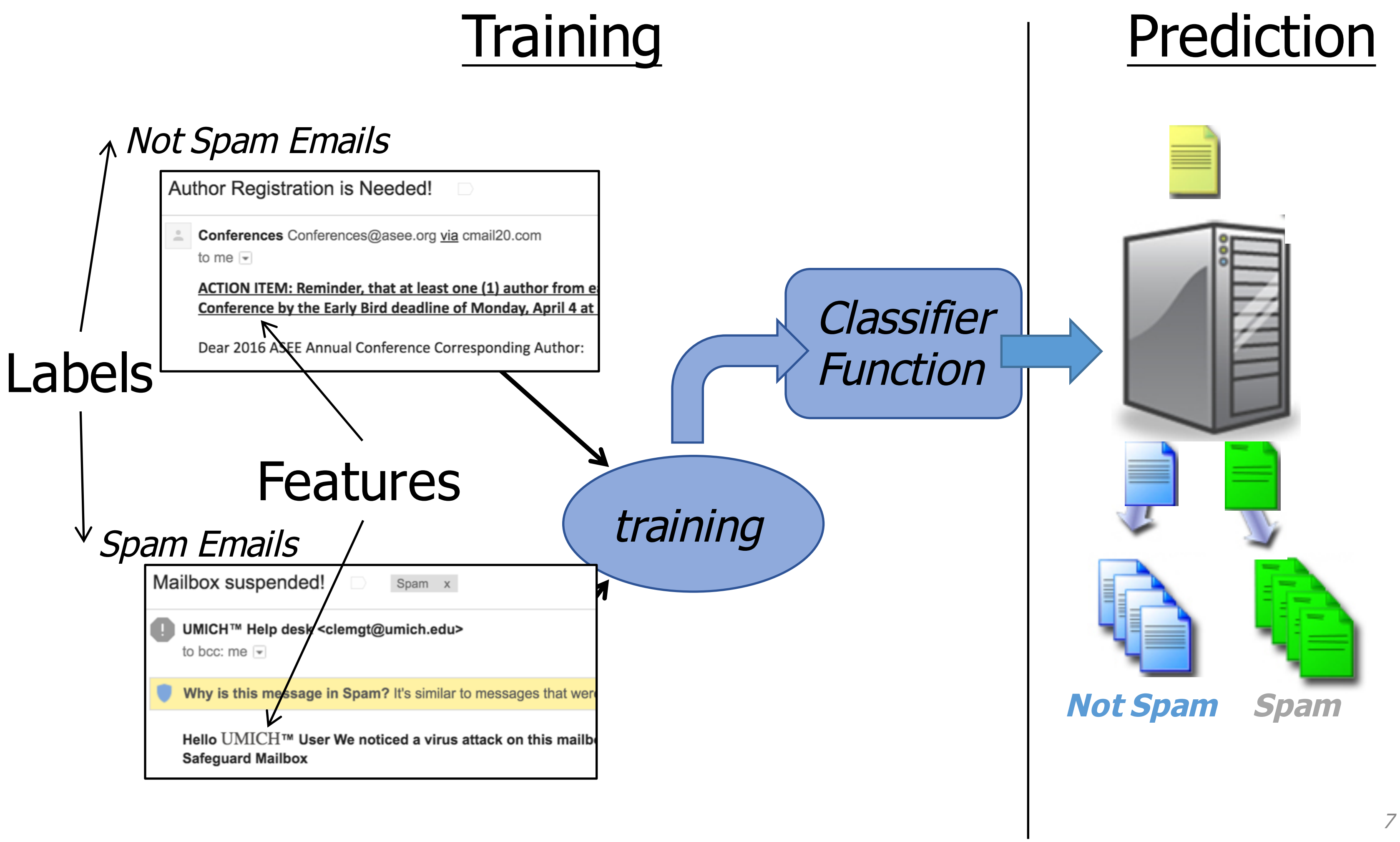
- Machine Learning is also known as data mining, or statistical methods
- Predates the Web, works fine on non-Web data
- But the Web throws off lots of easily-to-process human-activity data
  - A natural target for mining
  - Every successful Web company mines everything all the time
- What kinds of data have you generated today?

Machine Learning Introduction

- Informally, a learning algorithm is one that improves performance at a task with “experience”
- “Experience” == example data

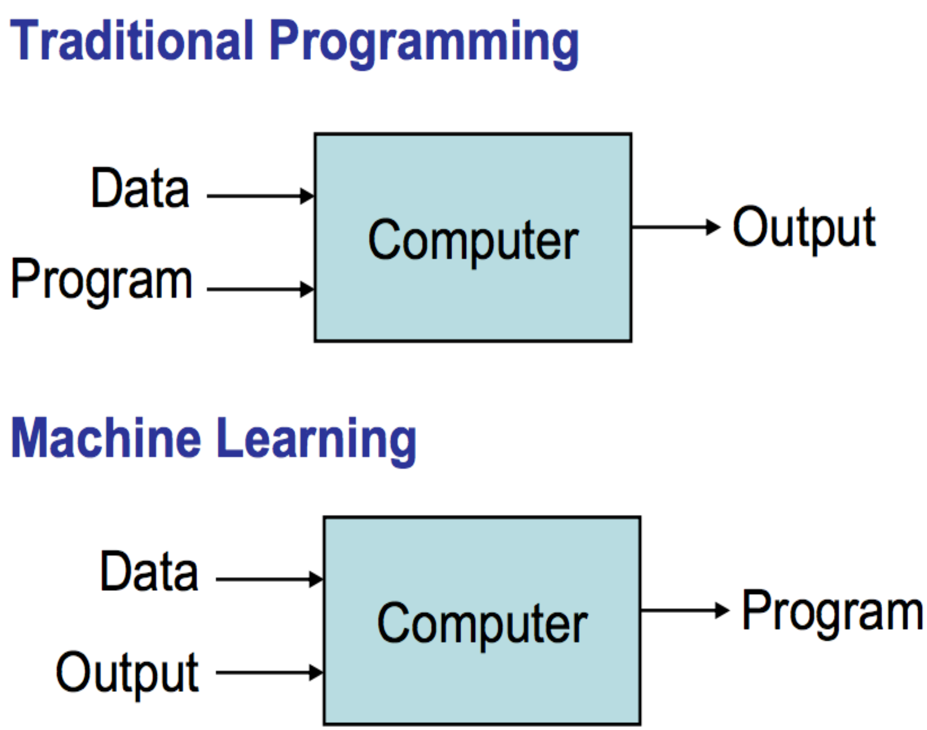
- Example: Spam filtering
  - Look at a bunch of emails that users have identified as "spam" and "not spam"
  - Predict whether a new email is "spam" or "not spam"

Example: Spam Filtering



Inputs and Outputs

- Informally, a learning algorithm is one that improves performance at a task with experience



Types of Learning

- Supervised learning
  - Example: Spam filtering
  - Training data with known correct answers
- Unsupervised learning
  - Example: grouping similar news articles
  - Training data without any answers
- Reinforcement learning
  - Example: robots that adapt to their environments
  - Training as-you-go

Types of Learning

Supervised learning examples:

- "Predict if this email is spam or not spam"
- "Predict the opening price for GOOG"
- "Predict likeliest next product purchase"

- *Training data contains example cases and the correct answer*
- Recommender systems use a form of supervised learning

Types of Learning

- Unsupervised learning
  - *Training data contains example cases, but not the correct answer*
  - "Show the natural clusters in the data"

- Reinforcement Learning
  - The learner gets a signal in the form of "good dog / bad dog"

Supervised Learning

- Inductive learning, or "prediction"
  - Given examples of a fn (X, F(X)) predict F(X) for a novel value X

- **Classification**
  - F(X) is discrete; is page relevant or not?
- **Regression**
  - F(X) is continuous; value of GOOG?
- **Probability estimation**
  - F(X) is probability of X; will Clinton win?

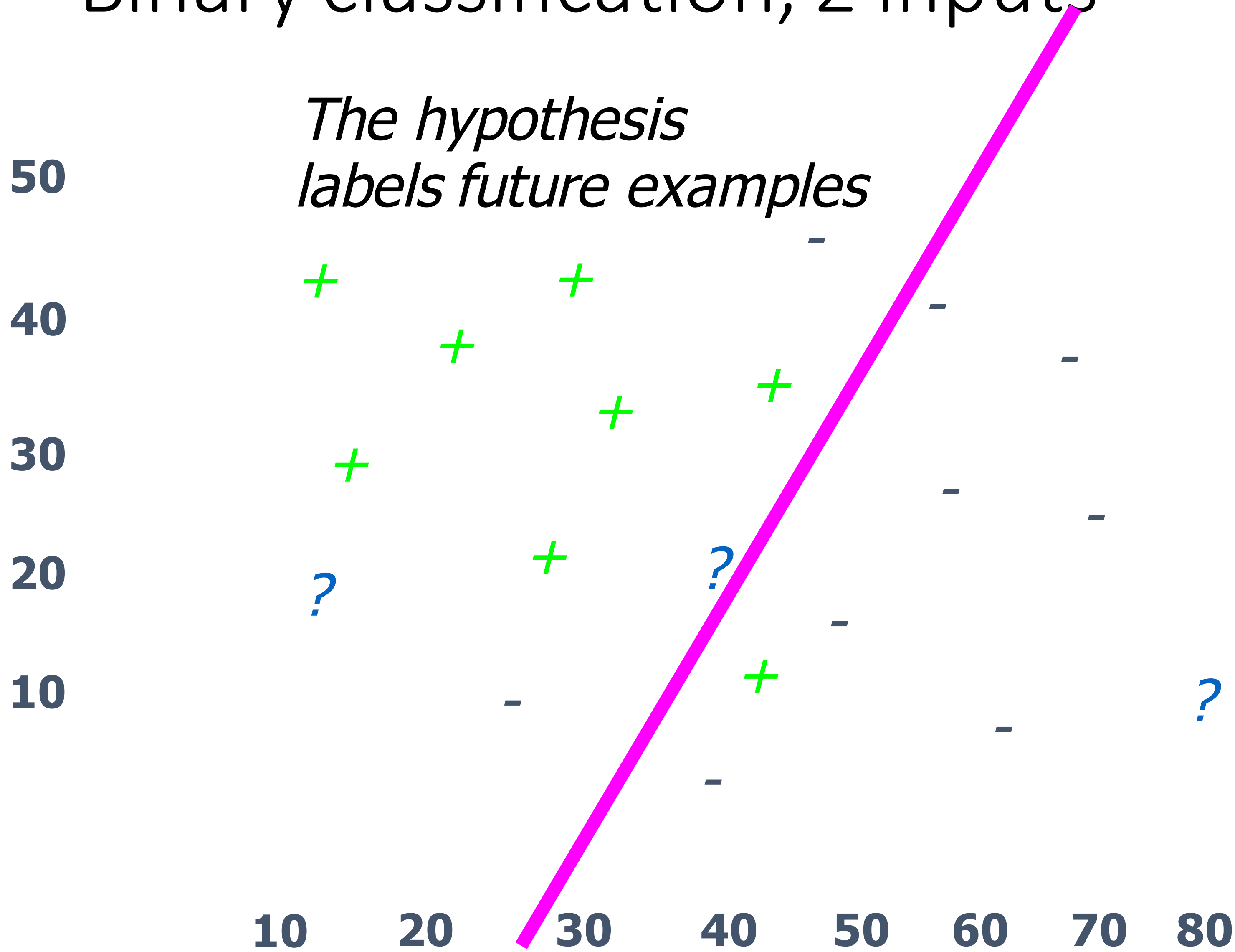
Supervised Learning

- Three ingredients
  - **Data**, with labels (e.g., the correct stock price)
  - **Features** (e.g., quarterly earnings, or num employees)
  - **Machine learning algorithm**

- Where do the labels come from?
- Where do the features come from?
- The set of labels tell the algorithm how to weigh evidence supplied by the features

- In many tasks, labeled data is in short supply and is the bottleneck

Binary classification; 2 inputs



Exercise

- Imagine you want to predict the age and gender of a Twitter user
  - What kind of supervised task is this?
  - What are the features?
  - What dataset is used to generate features?
  - How do you generate the labels?
- Imagine you want to classify a Jeopardy question as "geographic" or not
  - Same questions as above

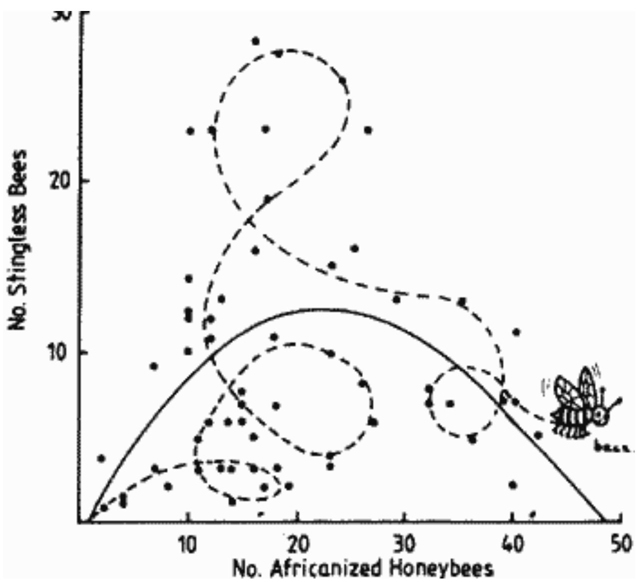
Discussion Question

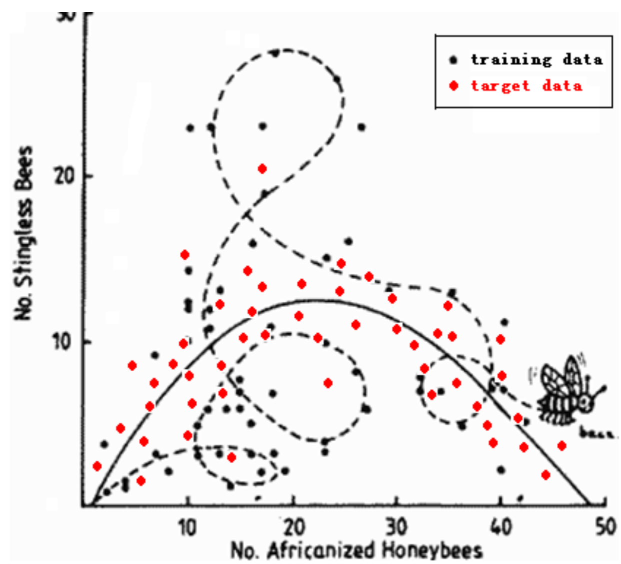
- Which one is best?
  - Perpendicular lines
  - Angled straight lines
  - Arbitrary curves

- True or false: "The best learner is the most flexible"

Overfitting

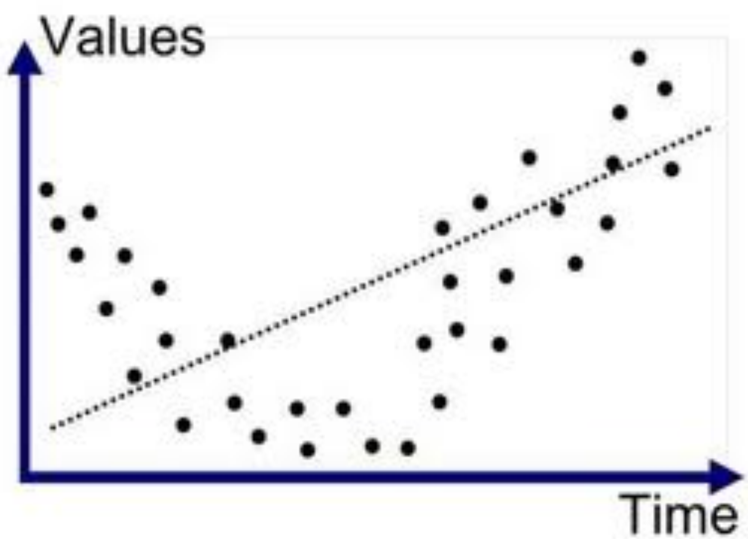
- Overfitting: model describes noise instead of the underlying relationship
  - Doesn't generalize well to new data
  - "Model is too complicated"





Underfitting

- Underfitting: model doesn't capture the underlying relationship
  - Doesn't generalize well to new data
  - "Model is too simple"



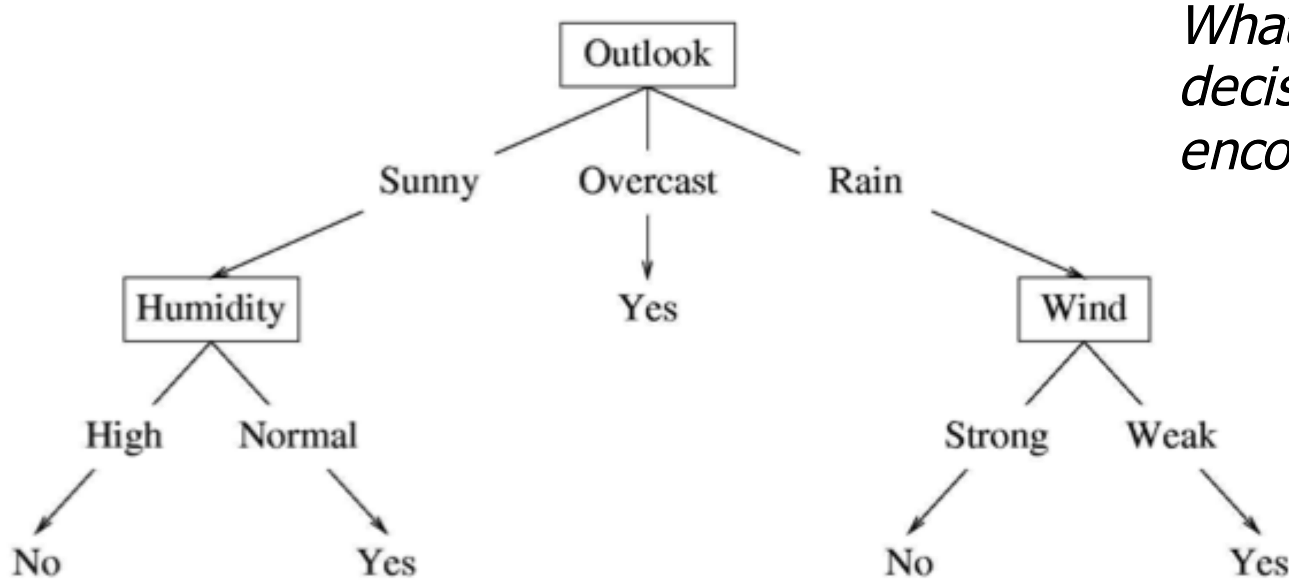
Performs poorly on both training data and new data

Bias

- Which hypotheses will be considered?
  - Lines?
  - Lines that are perpendicular to axes?
  - Circles?
  - Conic sections?
- Decision about the hypothesis space is called the *bias* of the machine learner
- Stronger (more restrictive) bias makes overfitting AND high accuracy harder

Classifiers

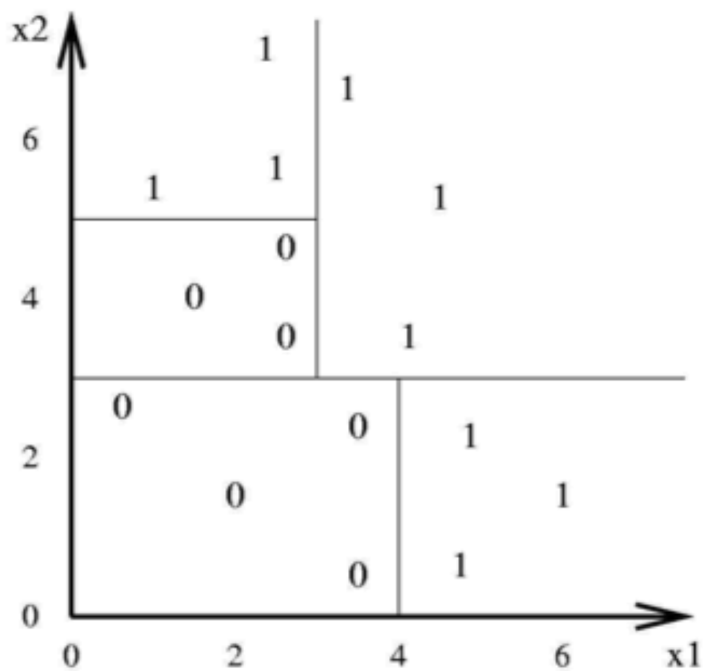
- Lots of classifier algorithms possible
- Decision Trees
  - Build a tree in which input variables are at internal nodes outputs at leaves

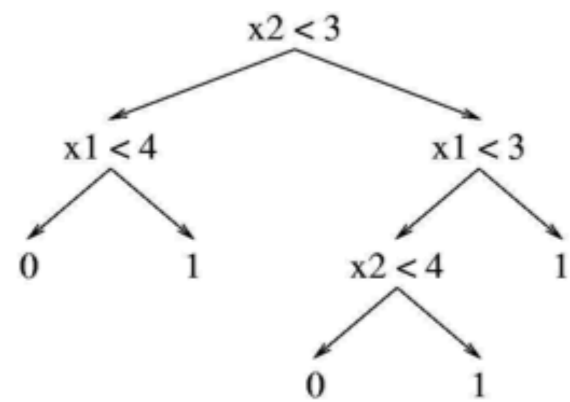


What is this decision tree encoding?

Classifiers

- Lots of classifier algorithms possible
- Decision Trees
  - Can express multiple decision regions, as long as they are parallel to axes



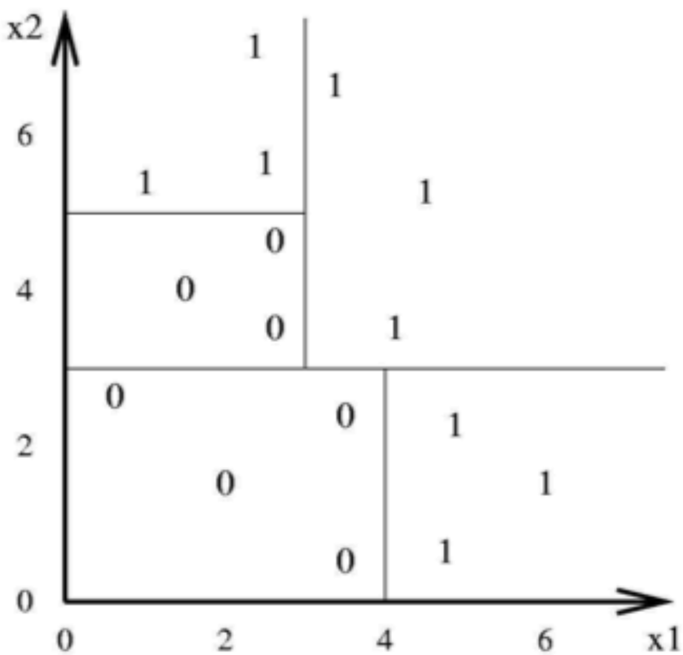


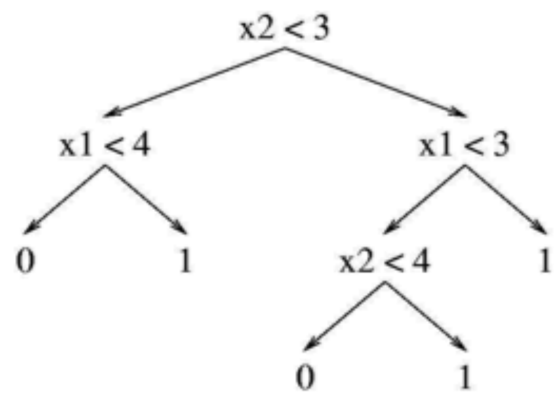
Classifiers

- Lots of classifier algorithms possible
- Rule Learners build a series of rules that are conjunctions of tests on input variables
  - Overcast => Yes
  - Sunny & Humid => Yes
  - Sunny & Normal => No
  - Rain & Strong-Rain => No
  - Rain & Weak-Rain => Yes
- Trees can always be converted to rules
- Vice-versa, as long as variables can appear multiple times in tree

In Detail: Decision Trees

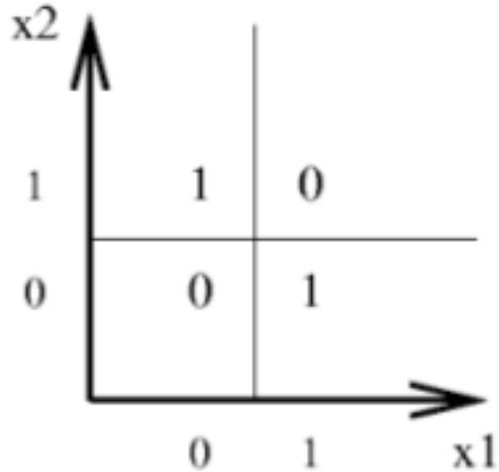
- Why are decision trees good?
  - Can represent any boolean function
  - Can handle discrete & continuous params
  - Easy for humans to understand, debug

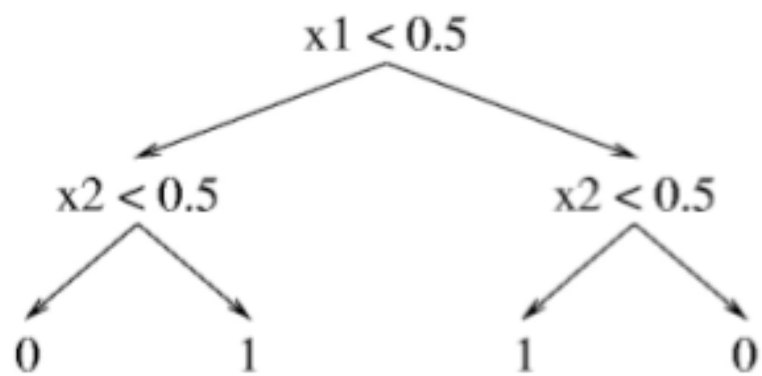




In Detail: Decision Trees

- But not necessarily an efficient representation of fn





DT Hypothesis Space

- As the number of nodes grows, the set of possible functions (the "hypothesis space") grows
  - Depth 1 ("decision stump"): 1 boolean input
  - Depth 2: any boolean fn of 2 inputs, and some of 3 $(x_1 \wedge x_2) \vee (\sim x_2 \wedge \sim x_3)$

Classification training set

- Training set S = {(**x**, y), …}
  - **x** is a vector of inputs
  - y is the desired output (label) for **x**
  - If **x** describes Outlook, Humidity, Wind then one value of **x** is [Sunny, High, Weak]
  - If y describes “whether or not to carry an umbrella”, then one value of y is “No”
- S is a set of (**x**, y) pairs
  - That is, many examples that describe the weather, plus whether or not to carry an umbrella

Experimental Evaluation

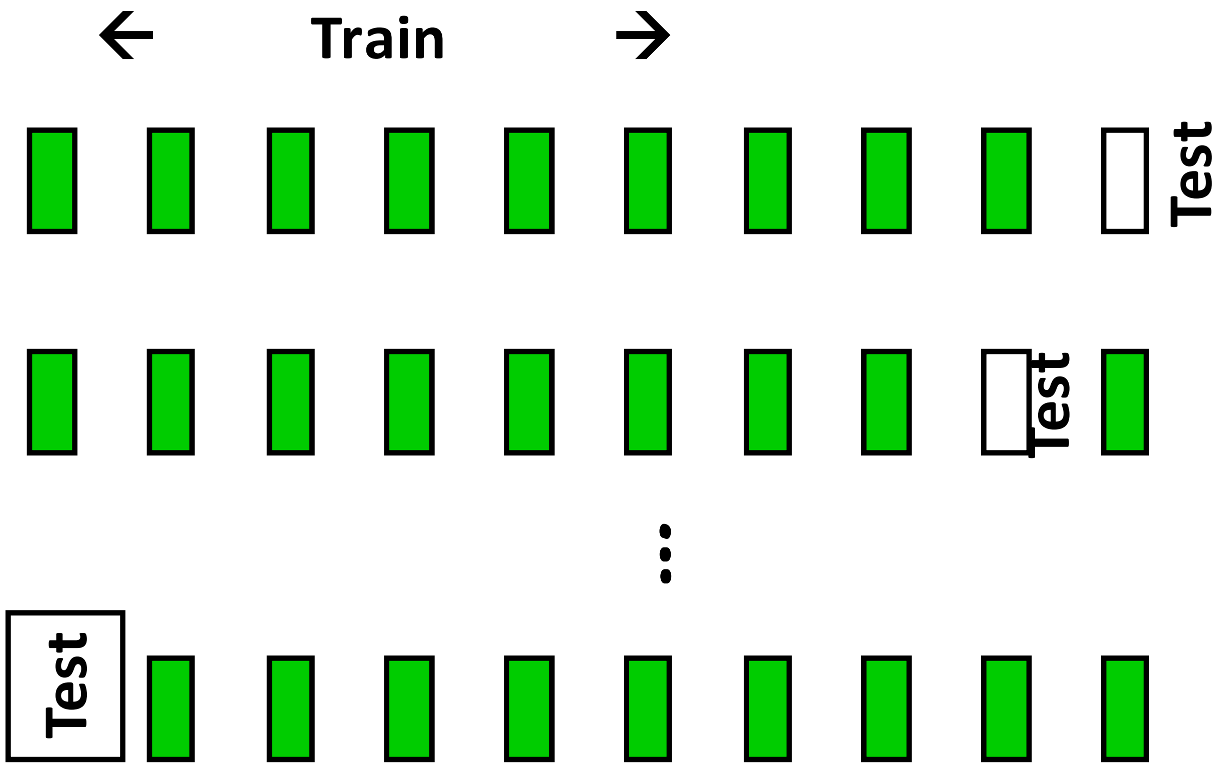
- How do we estimate the performance of classifier on unseen data?
- Can't just look at accuracy on training data – this will yield an over optimistic estimate of performance
- The test set must be held out during training
- Want to maximize training size, but still get accurate picture of performance
- Lots of data? Use 70/30 train/test split
- Performance == accuracy on test data

Experimental Evaluation

- What if you don't have much data?
  - Say, 10 data points?
  - More training data is better, but test set must be representative of future tasks

- Partition examples into *k* disjoint sets
- Now create *k* training sets
  - Each set is union of all equiv classes ***except one***
  - So each set has (k-1)/k of the original training data

Evaluation: Cross Validation



Cross-Validation

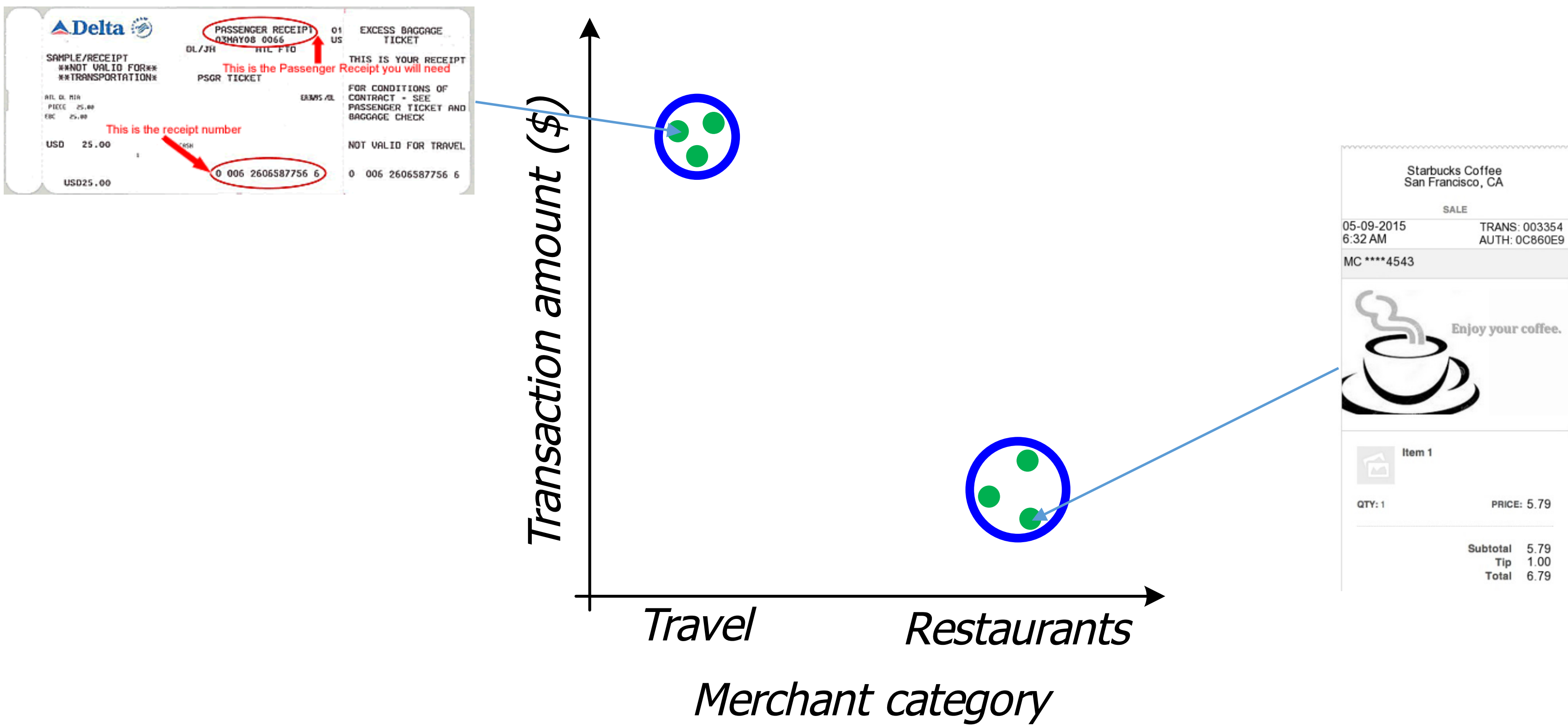
- Leave-one-out
  - Hold out one example, train on remaining
  - Train k learners; average the test results
  - Use if <~100 examples
- k-fold cross validation
  - Train k learners; use 1/k of data for test
  - If have ~100-~1000's of examples

Unsupervised Learning

- Sometimes called clustering
- Lots of applications in big data
  - Bioinformatics: group like genes
  - Web: page deduplication, friend management
  - Vision: recognize similar objects
- Can you learn the structure of the input data without any labels?
  - Group together things that are similar
  - Don’t group together things that are dissimilar

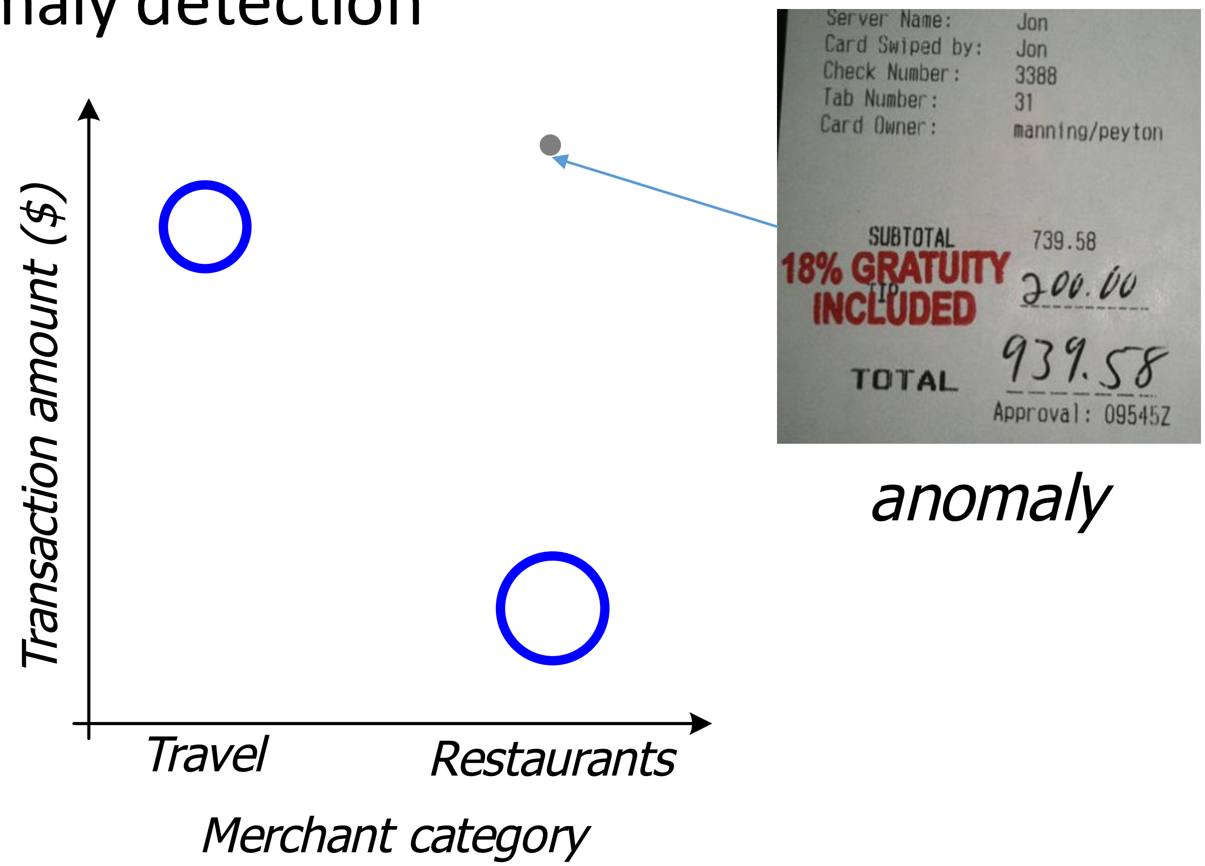
Unsupervised Learning Example

- Example: credit card fraud detection



Unsupervised Learning Example

- Good for anomaly detection



K-Means Clustering

- Very popular technique, as follows:
  - Grab a distance metric between two points
  - Choose the number of clusters = k
  - Generate k random "cluster centers"
- Repeat the following:
  - Assign each data item to the closest center
  - Choose new cluster centers
  - Until clusters don't move much
- In practice, distance metric matters more than clustering algorithm

When Does K-Means Fail?

- What if clusters are oblongs?
  - Rectangles?
  - Hourglasses?
- What if clusters overlap?
  - Document subsets?
  - Image closeups?
- What if clusters are different sizes?
  - People cloning wikipedia.org vs people cloning cafarella.com
  - Consider both volume and # points

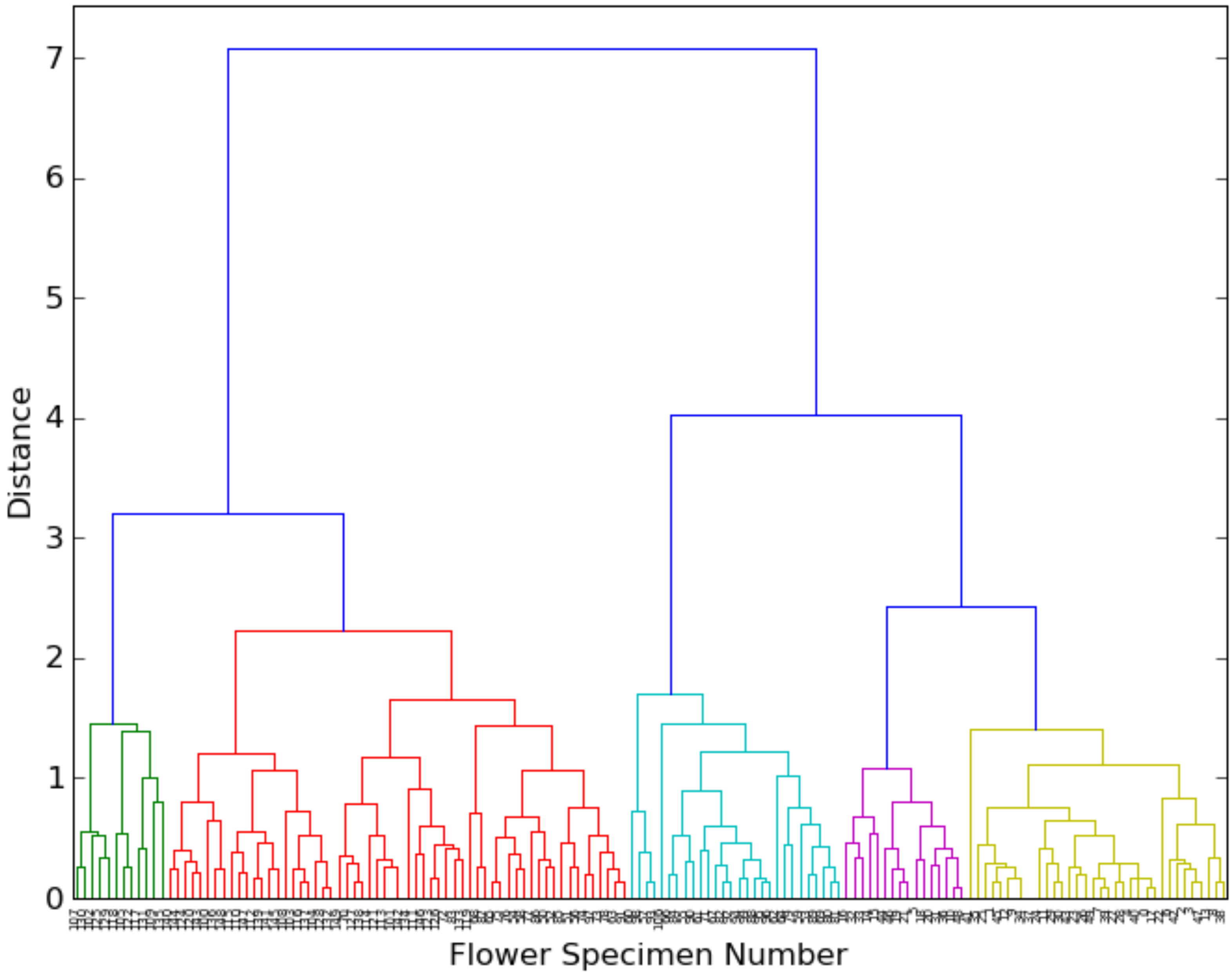
How To Pick k?

- Difficult without domain knowledge
- Agglomerative clustering
  - Start one cluster per example
  - Merge the two closest clusters
  - Repeat until you’ve got one cluster
  - Output result
- How do you measure cluster closeness?
  - Distance between centroids?
  - Min distance between pairs? (or max?)

Another way

- Divisive clustering
  - Put all samples into a single cluster
  - Split into two parts (via min-cut)
  - Repeat until you're happy with # of clusters

Sir Ronald Fisher's Iris Data Set



Cluster Evaluation

- Need the "right" number of "good" clusters
- Correctness of a cluster is easy
  - Do members belong together?
  - Roughly similar to precision
- Testing whether clusters are "right" is harder
- Multiple good clusterings possible for a single dataset
- In general, evaluation is much harder than with supervised learning

Important Questions

- How do you measure similarity?
- How do you construct the clustering?
- How do you evaluate the outcome?

Similarity Measurement

- Euclidean distance (for reals)
- Jaccard distance (for set overlap)
- Bit distance (for vectors of booleans)
- Normalized Mutual Information (NMI)
- Many others possible, depending on your application
- How would you measure similarity when clustering:
  - Images
  - Videos
  - Schemas

Ethics and Machine Learning

- Many data mining projects are ethically and politically contentious
  - Credit card offers
  - Financial trades
  - TIA project (Total Information Awareness)
- Many data-mining projects are ethically complicated because of the data used
  - Is the privacy-leaking AOL data OK?
  - What's so wrong about collecting WiFi info?