Machine Learning - Classification

Stanford University

Google

Google Cloud Platform

acm Association for Computing Machinery

Very Large Data Bases Endowment Inc.

INSTABASE

amazon web services™

Regression

Using data to make inferences or predictions

- Supervised
- Training data, each example:
  - Set of predictor values - "independent variables"
  - Numerical output value - "dependent variable"
- Model is function from predictors to output
  - Use model to predict output value for new predictor values
- Example
  - Predictors: mother height, father height, current age
  - Output: height

Classification

Using data to make inferences or predictions

- Supervised
- Training data, each example:
  - Set of feature values – numeric or categorical
  - Categorical output value - "label"
- Model is method from feature values to label
  - Use model to predict label for new feature values
- Example
  - Feature values: age, gender, income, profession
  - Label: buyer, non-buyer

Other Examples

Medical diagnosis

- Feature values: age, gender, history, symptom1-severity, symptom2-severity, test-result1, test-result2
- Label: disease

Email spam detection

- Feature values: sender-domain, length, #images, $keyword_1$, $keyword_2$, …, $keyword_n$
- Label: spam or not-spam

Credit card fraud detection

- Feature values: user, location, item, price
- Label: fraud or okay

Algorithms for Classification

Despite similarity of problem statement to regression, non-numerical nature of classification leads to completely different approaches

- K-nearest neighbors
- Decision trees
- Naïve Bayes
- … and others

K-Nearest Neighbors (KNN)

For any pair of data items i_1 and i_2, from their feature values compute *distance*(i_1,i_2)

Example:

Features - gender, profession, age, income, postal-code

person$_1$ = (male, teacher, 47, $25K, 94305)
person$_2$ = (female, teacher, 43, $28K, 94309)

distance(person$_1$, person$_2$)

distance() can be defined as inverse of *similarity*()

K-Nearest Neighbors (KNN)

Features - gender, profession, age, income, postal-code

$person_1$ = (male, teacher, 47, $25K, 94305)

$person_2$ = (female, teacher, 43, $28K, 94309)

Remember training data has labels

K-Nearest Neighbors (KNN)

Features - gender, profession, age, income, postal-code

$person_1$ = (male, teacher, 47, $25K, 94305) buyer

$person_2$ = (female, teacher, 43, $28K, 94309) non-buyer

Remember training data has labels

To classify a new item *i* : In the labeled data find the K closest items to *i*, assign most frequent label

$person_3$ = (female, doctor, 40, $40K, 95123)

KNN Example

- City temperatures – France and Germany
- Features: longitude, latitude
- Distance is Euclidean distance

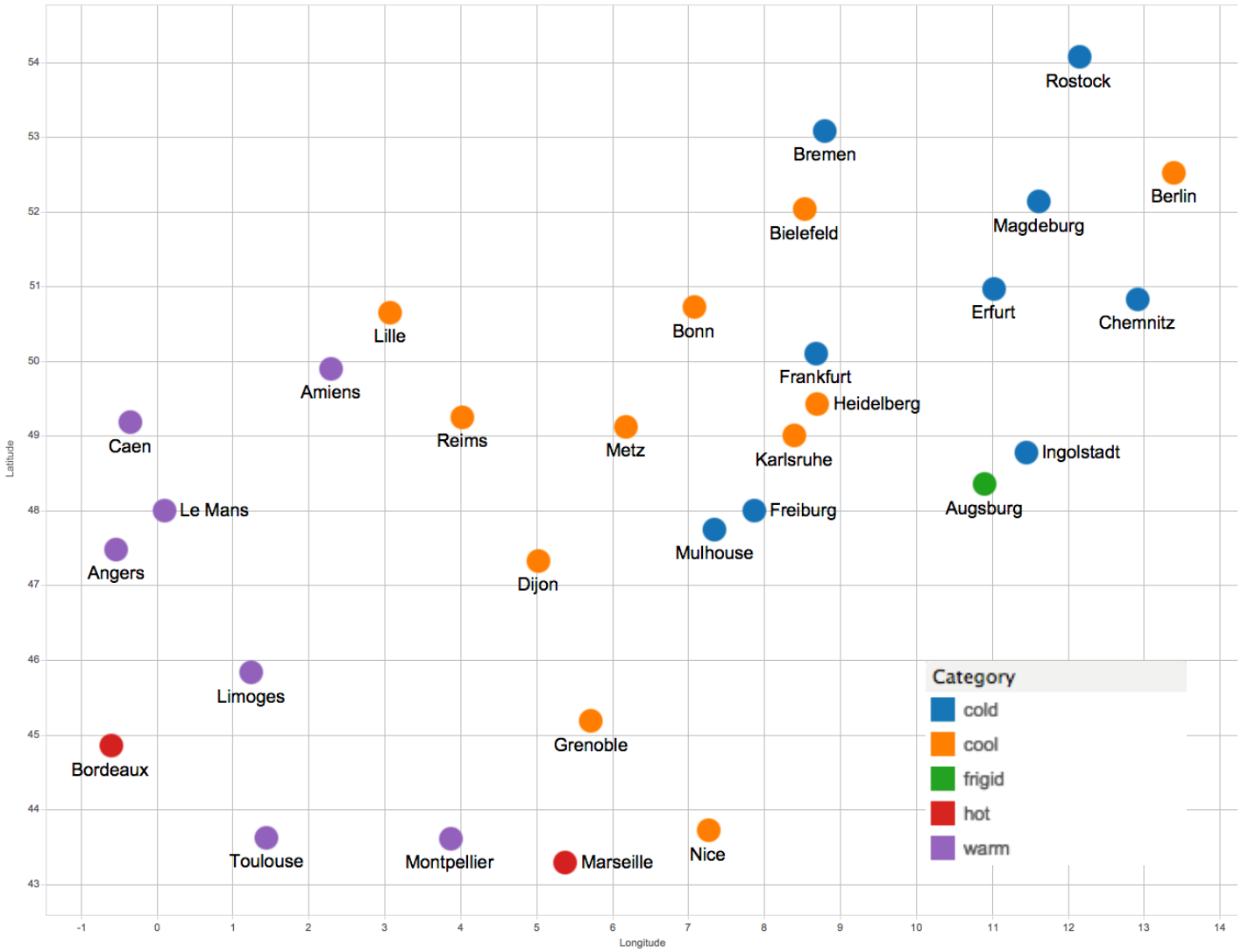
  $distance([o_1,a_1],[o_2,a_2]) = sqrt((o_1-o_2)^2 + (a_1-a_2)^2)$

  = actual distance in x-y plane
- Labels: frigid, cold, cool, warm, hot

Nice (7.27, 43.72) cool
Toulouse (1.45, 43.62) warm
Frankfurt (8.68, 50.1) cold
……

Predict temperature category from longitude and latitude

KNN Example


Rostock
Bremen
Berlin
Magdeburg
Bielefeld
Erfurt
Chemnitz
Lille
Bonn
Frankfurt
Amiens
Heidelberg
Caen
Reims
Metz
Karlsruhe
Ingolstadt
Augsburg
Le Mans
Freiburg
Mulhouse
Angers
Dijon
Limoges
Grenoble
Bordeaux
Nice
Toulouse
Montpellier
Marseille
Category
cold
cool
frigid
hot
warm
Latitude
Longitude
54
53
52
51
50
49
48
47
46
45
44
43
-1
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14

KNN Summary

To classify a new item *i*: find K closest items to *i* in the labeled data, assign most frequent label

- No hidden complicated math!
- Once distance function is defined, rest is easy
- Though not necessarily efficient

  Real examples often have thousands of features
  - Medical diagnosis: symptoms (yes/no), test results
  - Email spam detection: words (frequency)

  Database of labeled items might be enormous

"Regression" Using KNN

Features - gender, profession, age, income, postal-code

$person_1$ = (male, teacher, 47, $25K, 94305) buyer

$person_2$ = (female, teacher, 43, $28K, 94309) non-buyer

Remember training data has labels

To classify a new item *i*, find K closest items to *i* in the labeled data, assign most frequent label

$person_3$ = (female, doctor, 40, $40K, 95123)

"Regression" Using KNN

Features - gender, profession, age, income, postal-code

$person_1$ = (male, teacher, 47, $25K, 94305) $250

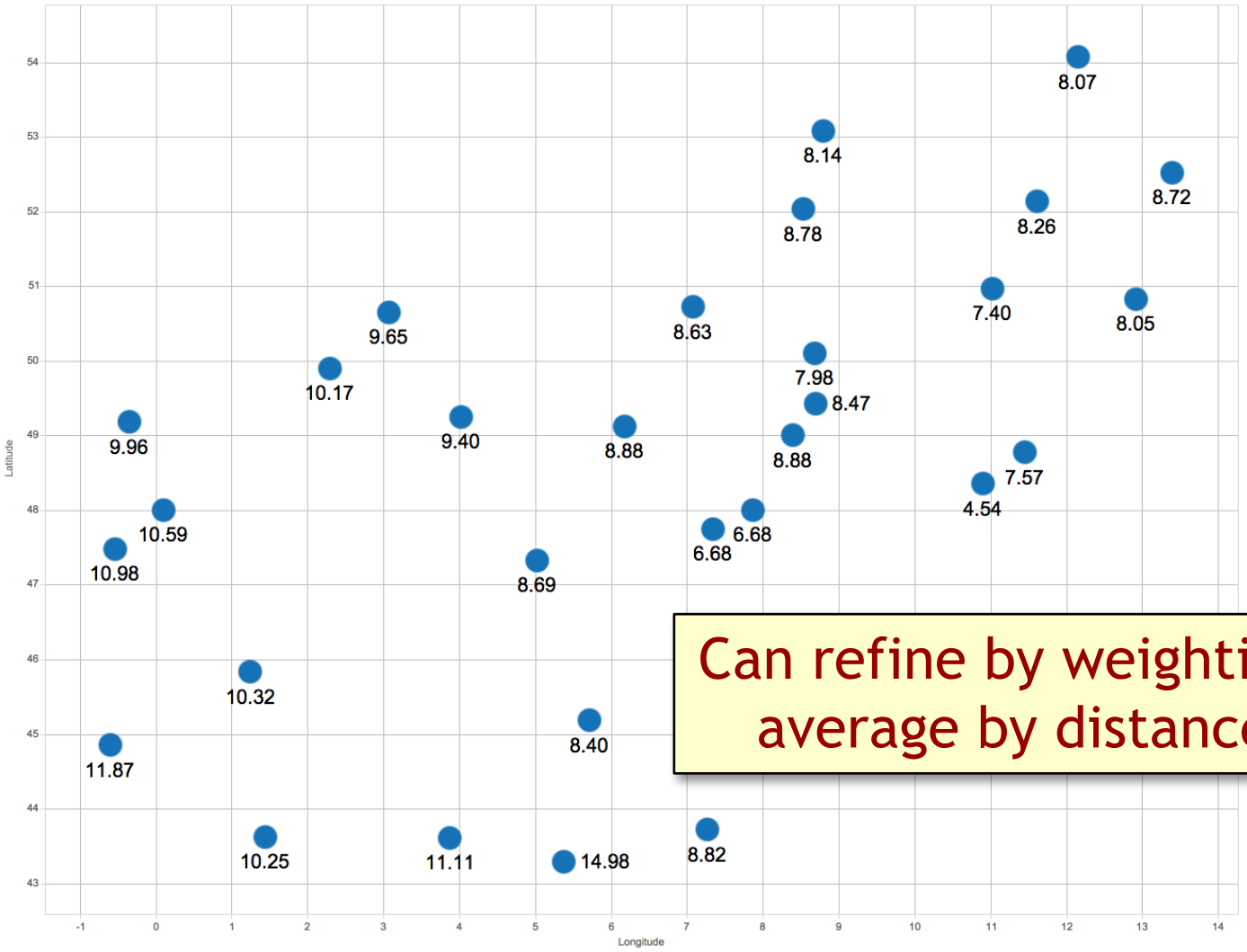
$person_2$ = (female, teacher, 43, $28K, 94309) $100

Remember training data has labels

To classify a new item *i*, find K closest items to *i* in the labeled data, assign average value of labels

$person_3$ = (female, doctor, 40, $40K, 95123)

Regression Using KNN - Example



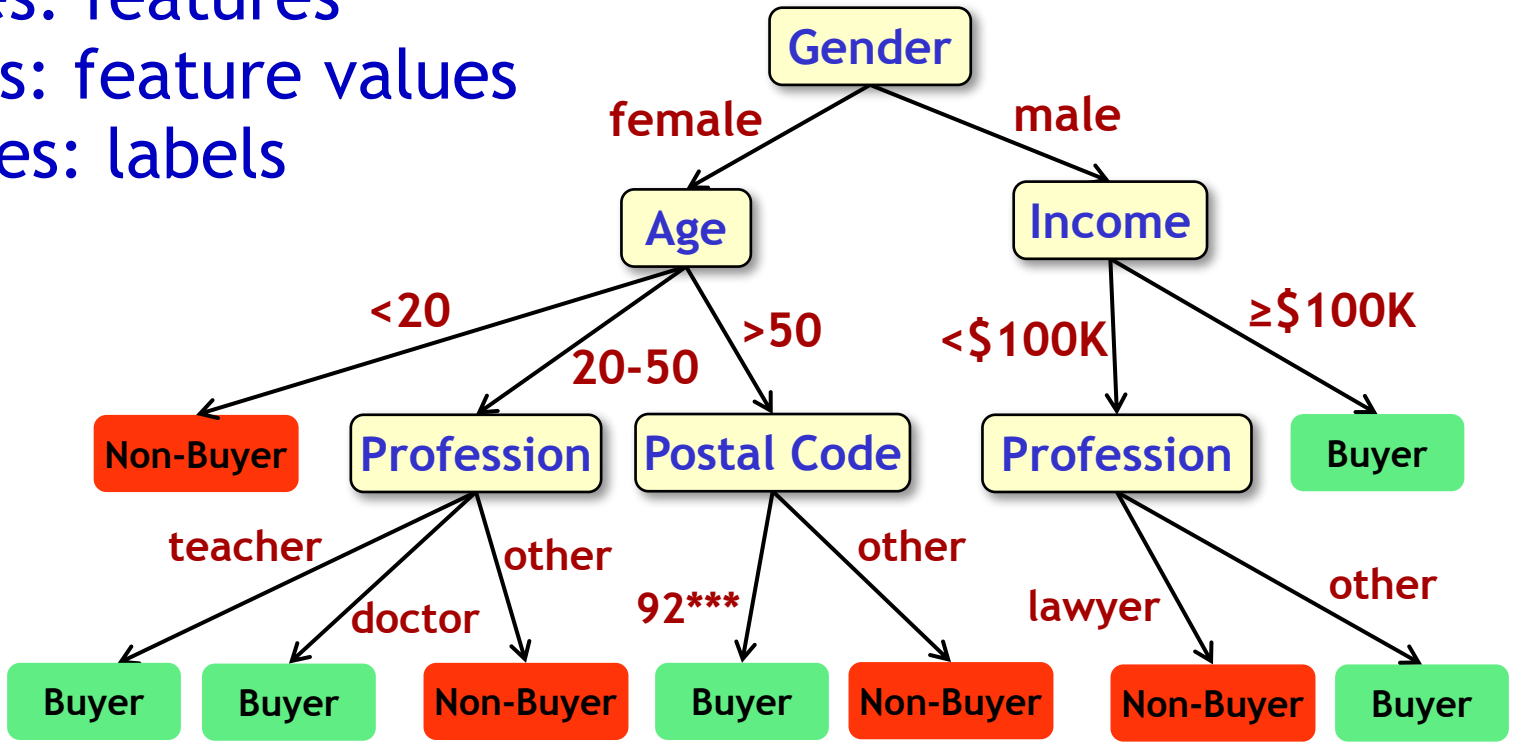
Can refine by weighting average by distance

Decision Trees

- Use the training data to construct a decision tree
- Use the decision tree to classify new data

Decision Trees

Nodes: features
Edges: feature values
Leaves: labels



New data item to classify:
Navigate tree based on feature values

Decision Trees

Primary challenge is building good decision trees from training data

- Which features and feature values to use at each choice point
- HUGE number of possible trees even with small number of features and values

Common approach: "forest" of many trees, combine the results

- Still impossible to consider all trees

Feature Selection

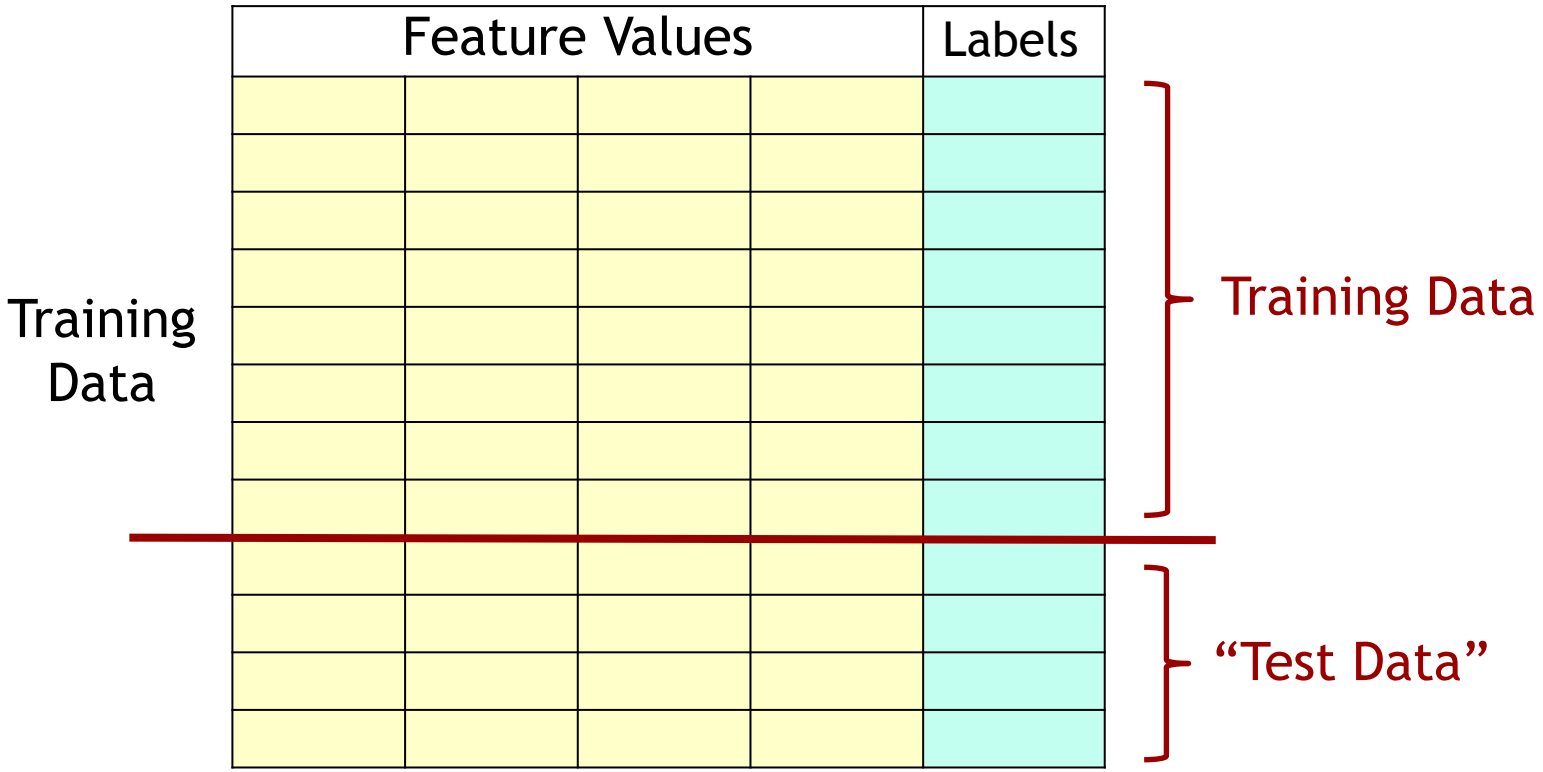
Real applications often have thousands of features

- Naïve Bayes typically uses only some of the features, those most affecting the label
- Decision trees also rely on choosing features that most affect the label
- Feature selection is a key part of machine learning – an art and a science

Training and Test

Created machine learning model from training data.
How do you know whether it's a good model?

- Try it on known data



Other Terms You Might Hear

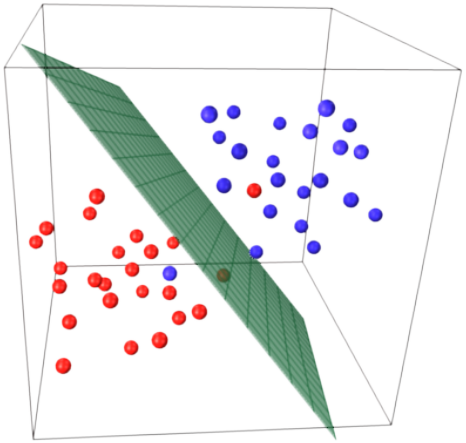
Logistic regression

- Recall regression model is function f from predictor values to numeric output value
- For classification: from training data obtain one regression function f_L for each label L

  $f_L(\textit{feature-values})$ = probability of item having label L

Support Vector Machine

- Two labels only ("binary classifier")
- Features = multidimensional space
- From training data SVM finds hyper-plane that best divides space according to labels



Other Terms You Might Hear

Deep Learning

- Complex, mysterious (the ultimate "black box" software), becoming extremely popular
- Multiple layers, each layer uses classification techniques to reduce complexity for next layer and further classification
- Important plus: identifies features from raw data

Neural Network

- Precursor to deep learning, typically two layers
- Leap to deep learning enabled by massive amounts of data, powerful computing

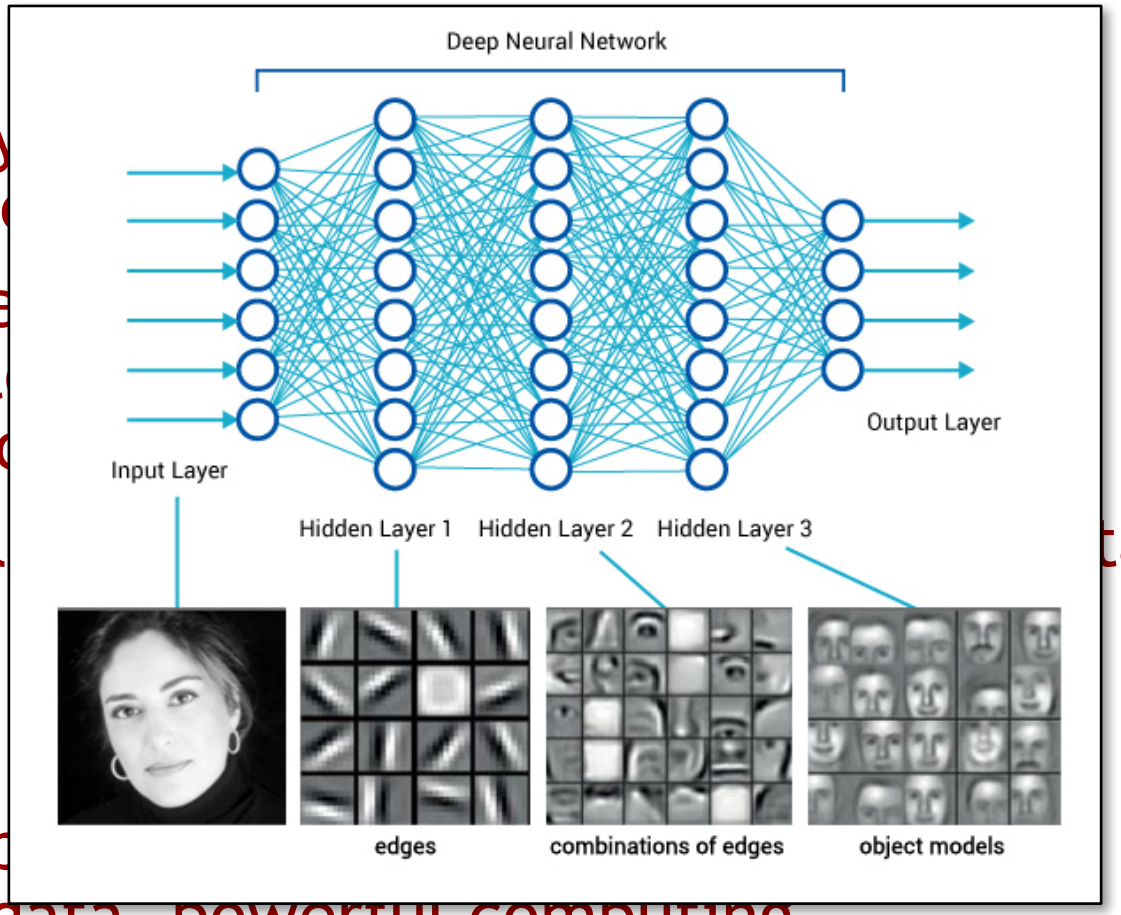
Other Terms You Might Hear

Deep Learning

- Complex, my
  software), b
- Multiple laye
  techniques t
  and further
- Important pl

Neural Network

- Precursor to
- Leap to deep
  amounts of data, powerful computing



Classification Summary

- Supervised machine learning
- Training data, each example:
  - Set of feature values – numeric or categorical
  - Categorical output value – label
- Model is "function" from feature values to label
  - Use model to predict label for new feature values
- Approaches we covered
  - K-nearest neighbors – relies on distance (or similarity) function
  - Decision trees – relies on finding good trees/forests
  - Naïve Bayes – relies on conditional independence assumption