# Lecture Notes for
# Machine Learning in Python

Professor Eric Larson

Week One, Lecture One

# Class Logistics and Agenda

- Syllabus
- Data Mining and Machine Learning
- Types of Data and Data Categorization

synchronous

# Course Syllabus

# Introductions

- Me
  - Eric Larson
- You
  - Name, department, grad/ugrad
  - Something true or false
- My approach to this course
  - programming
  - math
  - **applications** and **analytics**

synchronous

# The course syllabus

- Text: None
  - Recommended: Python Machine Learning, Sebastian Raschka
- Use Canvas for posted course material
- **Prerequisite**: ability to learn quickly these topics
  - Linear Algebra, Calculus
  - Basic statistics and probability
  - Python programming
- Grading:
  - Lab Assignments: 75% of grade (3 labs @ 25% each)
  - In Class: 20% of grade (4 at 5% each)
  - In Class Participation: 5% (yes, actually graded)

synchronous

# How will you grade participation

- Choose to respond to the question:

- Do you think this will work?

- A: Yes this is going to work
- B: This is not going to work:
- C: Wait, what…

Question ML1a

# Lab Assignments

- Lab assignments will be submitted electronically. Late labs will not be accepted.
- Lab assignments must be completed as a team.
- Lab assignments should be turned in as rendered jupyter notebook
- Most assignments are turned in during a week **where formal lecture does not take place:** use this extra time to complete time consuming analyses of the data
- There is a high expectation for these assignments. Comment code and explain reasoning in detail

synchronous

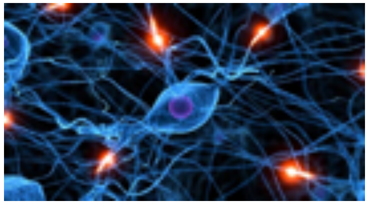
# Grading Rubric

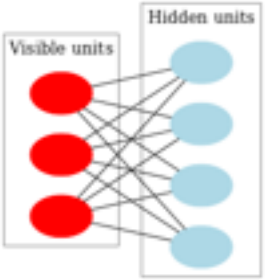
- In all assignments specific deliverables are asked and should be completed to the best of your ability.
- Each deliverable will be worth a certain percentage of the lab grade and you will be graded in terms of the quality of your analysis.
- Markup code so that it is readable and **immediately** understandable.
- The sum total of the these deliverables will be 90% of the points possible for each assignment. If you complete all the project deliverables satisfactorily you should expect a grade of 90%.
- The remaining 10% of the points are reserved for exceptional work and/or work that is above and beyond in one or more elements of the analysis.

# Machine Learning and Data Mining

# A History of Machine Learning

- Historically builds from disciplines statistics and computer science (algorithms)
- Its really just algorithms for learning

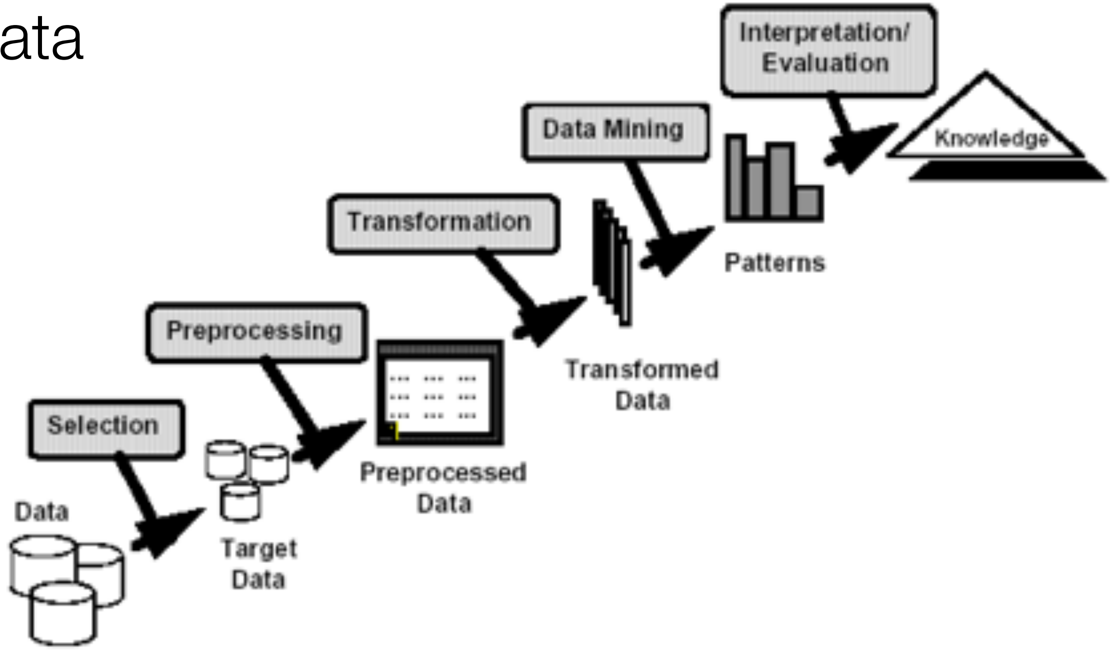




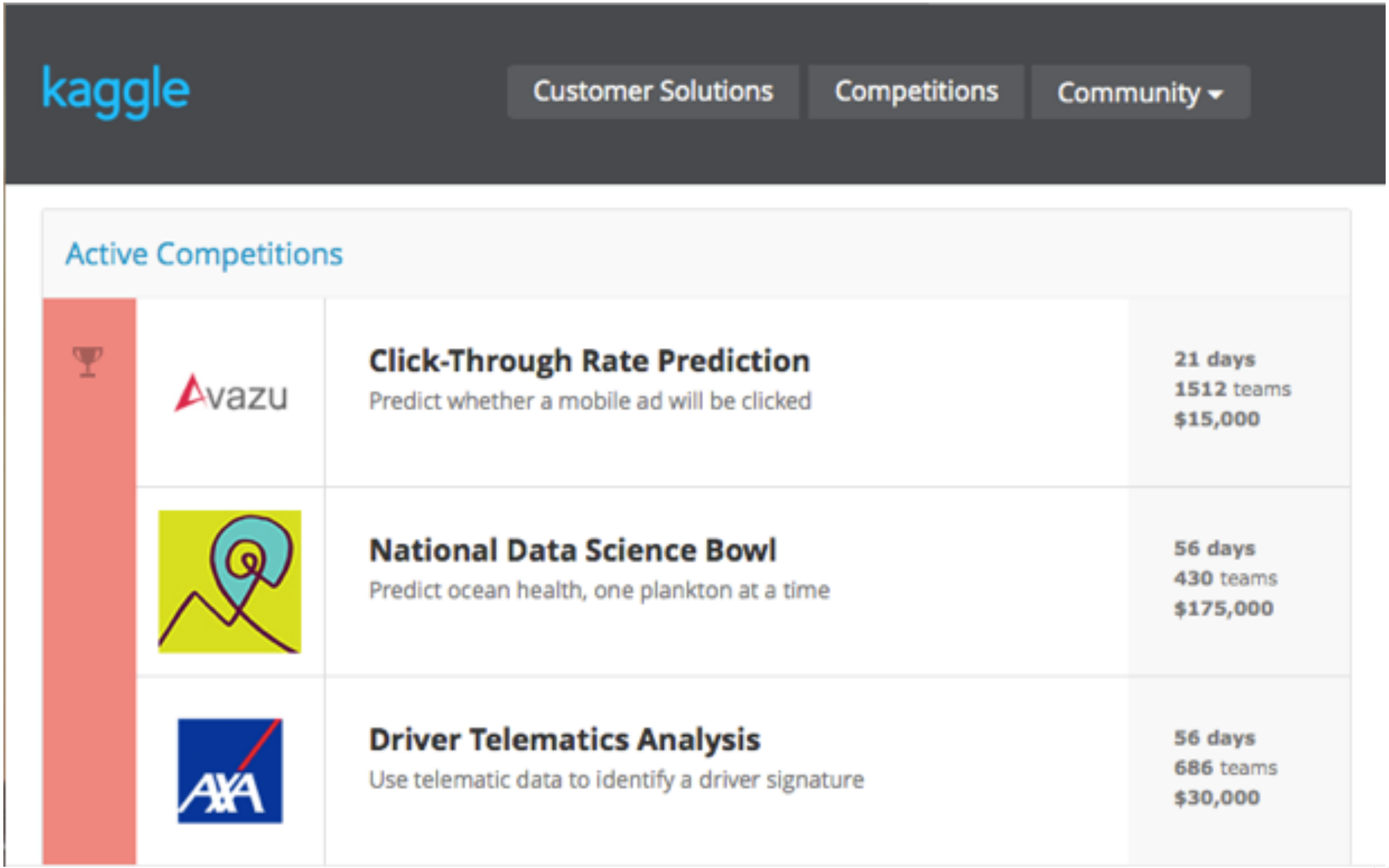
- 1952: Arthur Samuel IBM creates checker program
- 1957: Rosenblatt, Neural Network Perceptron
- 1967: Nearest Neighbor Pattern Recognition
- 1970's: AI Winter
- 1990's: Volley of new Machine learning Algorithms
- 2001: Breiman's Random Forests
- ~2004: Modern Support Vector Machines with Kernels
- ~2010: Deep Learning Convolutional Networks
- 2015: Deep Learning becomes buzz word, you hear about it and take this course for 2016

# What is Machine Learning?

- Many Definitions
  - Non-trivial extraction of **implicit**, previously **unknown**, and potentially **useful** information from data
  - Exploration & analysis, by **automatic** or **semi-automatic** means over large quantities of data in order to discover meaningful patterns



# Contemporary problems in Machine Learning



Active Competitions

| | | |
|---|---|---|
| Avazu | **Click-Through Rate Prediction**<br>Predict whether a mobile ad will be clicked | 21 days<br>1512 teams<br>$15,000 |
| | **National Data Science Bowl**<br>Predict ocean health, one plankton at a time | 56 days<br>430 teams<br>$175,000 |
| AXA | **Driver Telematics Analysis**<br>Use telematic data to identify a driver signature | 56 days<br>686 teams<br>$30,000 |

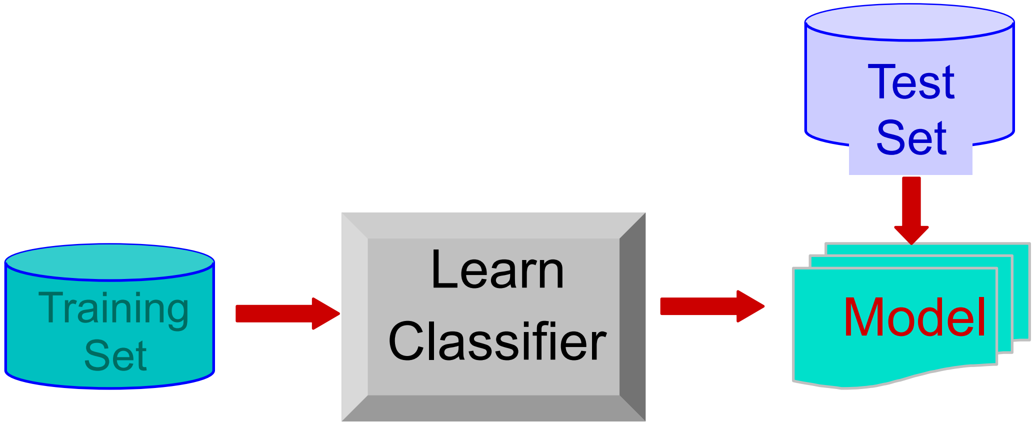
# Data Mining and Machine Learning

- Prediction Methods
  - Use some variables to predict unknown or future values of other variables
- Description Methods
  - Find human-interpretable patterns that describe the data.

- Classification [Predictive]
- Regression [Predictive]
- Deviation Detection [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]

# Classification: Definition

- Given a collection of records (*training set* )
  - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Find a *model* for class attribute as a function of the values of other attributes.
- Goal: <u>previously unseen</u> records should be assigned a class as accurately as possible.



# Classification: Definition

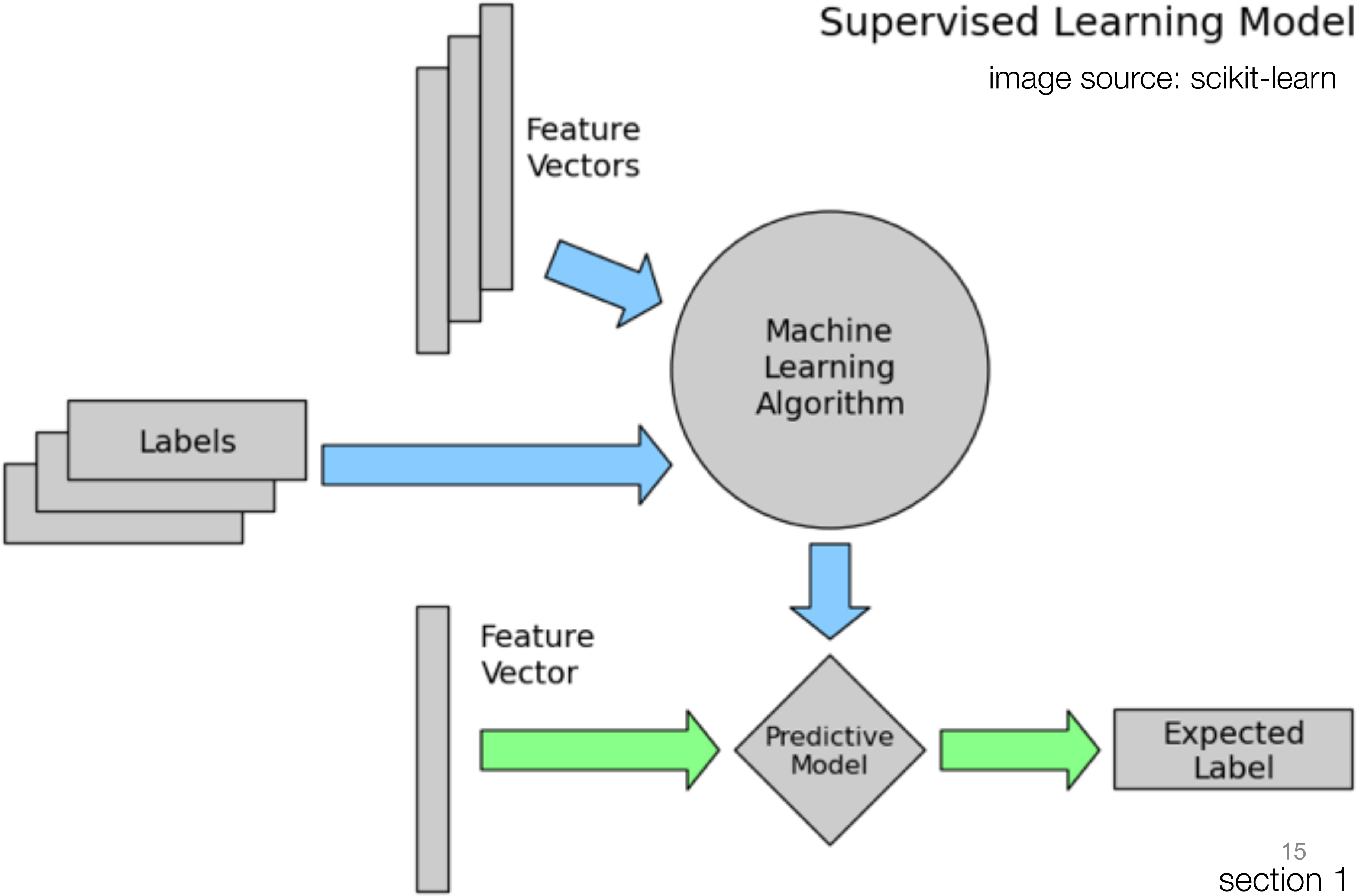


image source: scikit-learn

# Classification: Application 1

- Direct Marketing
  - Goal: Reduce cost of mailing by *targeting* a set of consumers likely to buy a new cell-phone product.
  - Approach:
    - Use the data for a similar product introduced before.
    - *{buy, don't buy}* decision forms the *class attribute*.
    - Collect various demographic, lifestyle, and company-interaction related information about all such customers.

Training Set

| TID | Job | Earning | Class |
|---|---|---|---|
| 1 | Lawyer | $310k | Buy |
| 2 | Doctor | $265k | Don't Buy |
| 3 | Student | $20k | Buy |
| 4 | Prof. | $1M | Buy |

Unknown

| ***TID*** | ***Job*** | ***Earning*** |
|---|---|---|
| ***1*** | Student | $3k |

From [Berry & Linoff] Data Mining Techniques, 1997

# Classification: Application 2

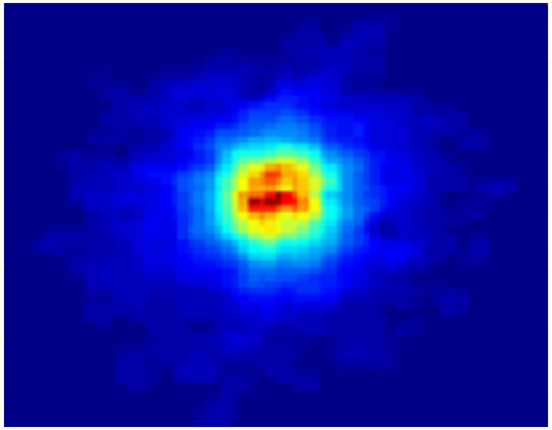
- Sky Survey Cataloging
  - Goal: To predict class (star or galaxy) of sky objects, especially visually faint ones, based on the telescopic survey images (from Palomar Observatory).
    - 3000 images with 23,040 x 23,040 pixels per image.
  - Approach:
    - Segment the image.
    - Measure image attributes (features) - 40 of them per object.
    - Model the class based on these features.
    - Success Story: Could find 16 new high red-shift quasars, some of the farthest objects that are difficult to find!

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

# Classifying Galaxies

Courtesy: http://aps.umn.edu

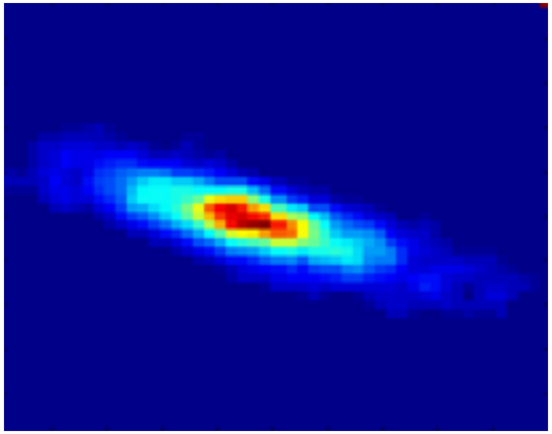
*Early*



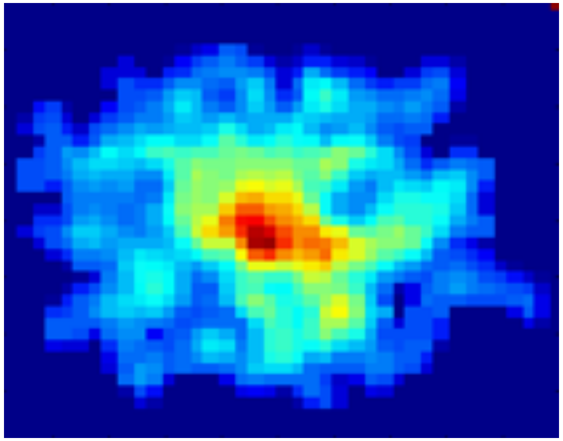
Class:
- Stages of Formation

Attributes:
- Image features,
- Characteristics of light waves received, etc.

*Intermediate*



*Late*



Data Size:
- 72 million stars, 20 million galaxies
- Object Catalog: 9 GB
- Image Database: 150 GB

# Regression

- Predict a value of a given *continuous valued* variable based on the values of other variables
- Examples:
  - Predicting sales amounts of new product based on advertising expenditure.
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
  - Predicting lung function as a function of gender, weight, height

**Training Set**

| *TI* | *Gende* | *Weight* | *Asthma* | *LF* |
|---|---|---|---|---|
| *1* | *M* | *175lbs* | *N* | *85%* |
| *2* | *F* | *150lbs* | *N* | *87.3%* |
| *3* | *F* | *155lbs* | *Y* | *90%* |
| *4* | *M* | *225lbs* | *Y* | *65.2%* |

**Unknown**

| *TI* | *Gende* | *Weight* | *Asthma* |
|---|---|---|---|
| *1* | *M* | *160lbs* | *N* |

# Self Test

- (**A. classification)**
  **(B. regression)**
  **(C. not Machine Learning)**
  - Dividing up customers by potential profitability?
    - classification/regression
  - Extracting frequency of sound?
    - NOT ML
  - Finding someone's adipose tissue measure from waist circumference?
    - regression
  - Deciding if a person has diabetes based upon their history and diet?
    - classification
  - Finding the genre of an online article based on the words in it?
    - classification

# Types of Data and Categorization

# What is Data?

- Collection of data **objects** and their **attributes**

- An **attribute** is a property or characteristic of an object
  - Examples: eye color of a person, temperature, etc.

- A collection of attributes describe an **object**

**Attributes**, variables, fields, characteristics, features

**Objects**, records, points, samples, cases, entities, instances

| *TID* | *Pregnant* | *BMI* | *Age* | *Diabetes* |
|---|---|---|---|---|
| *1* | Y | 33.6 | 41-50 | positive |
| *2* | N | 26.6 | 31-40 | negative |
| *3* | Y | 23.3 | 31-40 | positive |
| *4* | N | 28.1 | 21-30 | negative |
| *5* | N | 43.1 | 31-40 | positive |
| *6* | Y | 25.6 | 21-30 | negative |
| *7* | Y | 31.0 | 21-30 | positive |
| *8* | Y | 35.3 | 21-30 | negative |
| *9* | N | 30.5 | 51-60 | positive |
| *10* | Y | 37.6 | 51-60 | positive |

# Types of Attributes

- There are different types of attributes
  - **Nominal**
    - Examples: ID numbers, eye color, zip codes
  - **Ordinal**
    - Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height in {tall, medium, short}
  - **Interval**
    - Examples: calendar dates, temperatures in Celsius or Fahrenheit.
  - **Ratio**
    - Examples: temperature in Kelvin, length, time, counts

# Properties of Attribute Values

- The type of an attribute depends on which of the following properties it possesses:
    - Distinctness: = ≠
    - Order: < >
    - Addition: + -
    - Multiplication: * /

    - **Nominal** attribute: distinctness
    - **Ordinal** attribute: distinctness & order
    - **Interval** attribute: distinctness, order, & addition
    - **Ratio** attribute:distinctness, order, addition, multiplication

| Attribute Type | Description | Examples | Operations |
|---|---|---|---|
| **Nominal** | The values are different names, i.e., only enough information to distinguish one object from another. (=, ≠) | zip codes, employee ID numbers, eye color, sex: {male, female} | mode, entropy, contingency correlation, $\chi^2$ test |
| **Ordinal** | The values of an ordinal attribute provide enough information to order objects. (<, >) | hardness of minerals, {good, better, best}, grades, street numbers | median, percentiles, rank correlation, run tests, sign tests |
| **Interval** | For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. (+, - ) | calendar dates, temperature in Celsius or Fahrenheit | mean, standard deviation, Pearson's correlation, t and F tests |
| **Ratio** | For ratio variables, both differences and ratios are meaningful. (*, /) | temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current | geometric mean, harmonic mean, percent variation |

# Feature Type Representation

| | Attribute | Representation Transformation | Comments |
|---|---|---|---|
| Discrete | Nominal | Any permutation of values<br>one hot encoding | If all employee ID numbers were reassigned, would it make any difference? |
| | Ordinal | An order preserving change of values, i.e., new_value = f(old_value) where f is a monotonic function.<br>integer | An attribute encompassing the notion of good, better best can be represented equally well by the values {1, 2, 3} or by { 0.5, 1, 10}. |
| Continuous | Interval | new_value =a * old_value + b where a and b are constants<br>float | Thus, the Fahrenheit and Celsius temperature scales differ in terms of where their zero value is and the size of a unit (degree). |
| | Ratio | new_value = a * old_value<br>float | Length can be measured in meters or feet. |

# Self Test

- Are these **A. interval or B. ratio**:
    - Angle measured 0-360 degrees
        - ratio
    - Height above sea level
        - interval or ratio depending on if sea level is considered arbitrary
- Are these **A. ordinal, B. nominal, or C. binary**?
    - military rank
        - ordinal
    - coat check number
        - nominal
    - time as AM or PM
        - binary

# Before Next Lecture

- Before next class:
  - install python on your laptop
  - install anaconda distribution of python

- Look at Python primer if you need an intro to Python

synchronous

# Demo

If time:
Jupyter Notebooks
and Numpy

synchronous