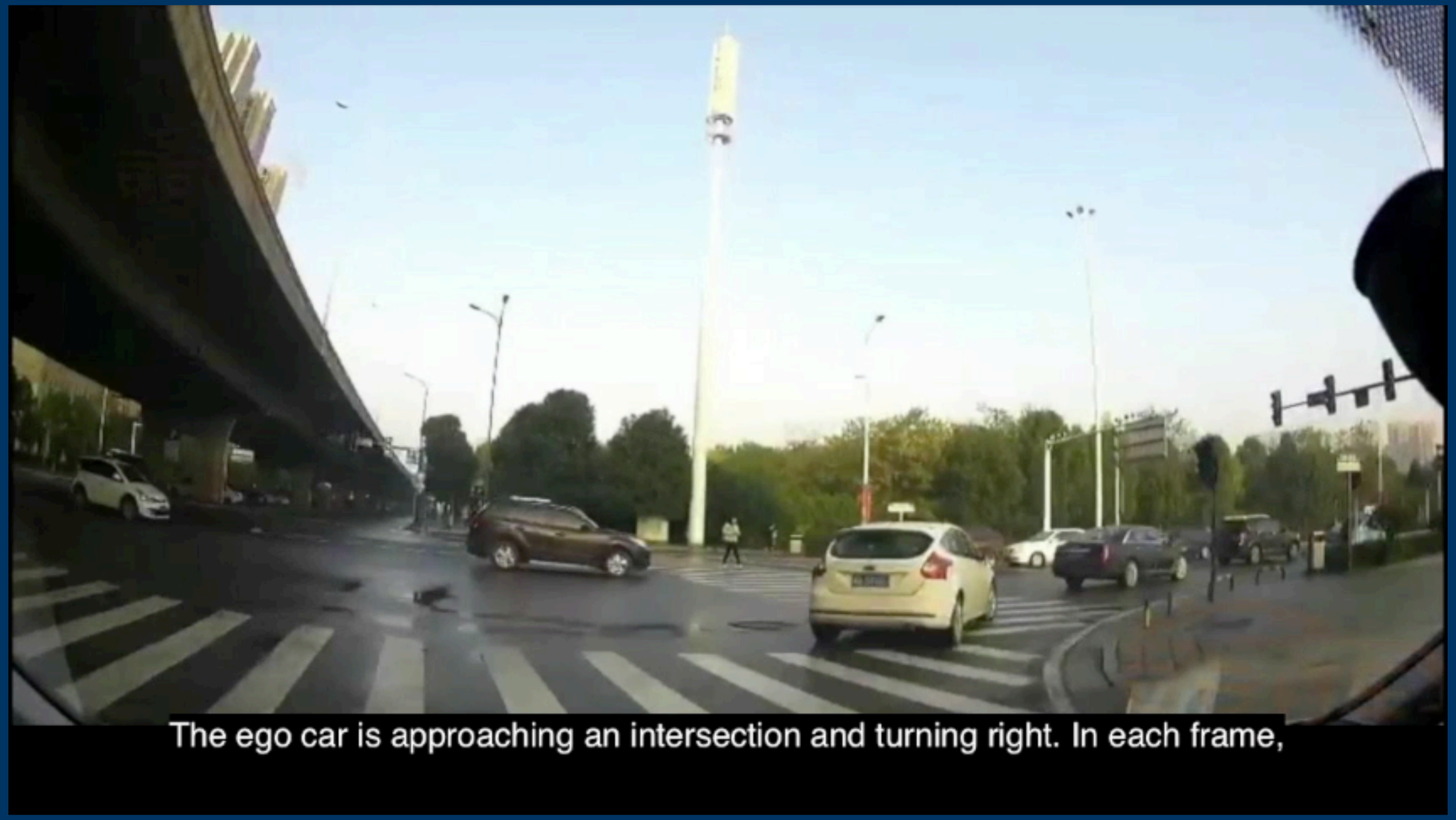
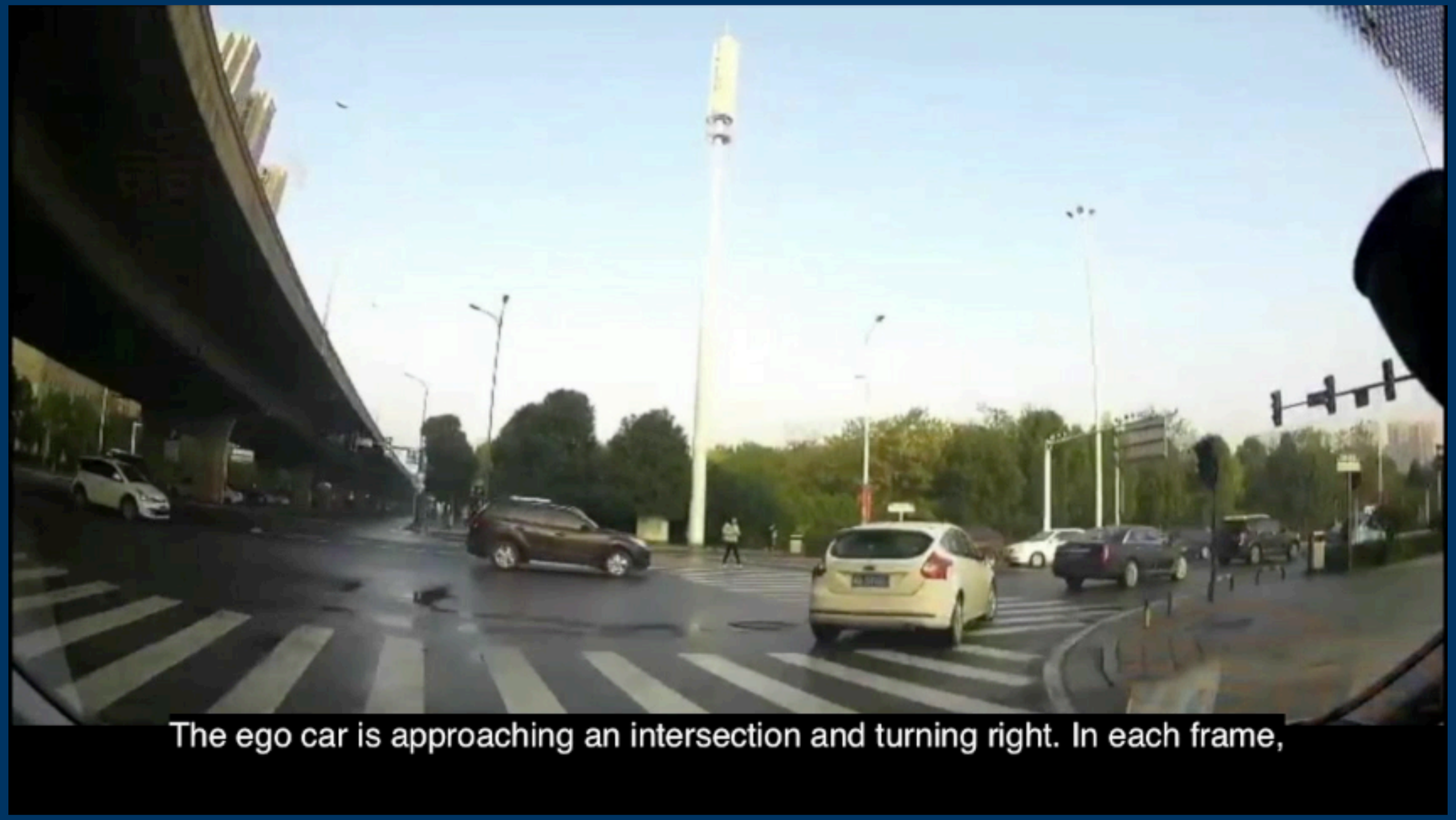


Basics of Machine Learning

CSCI-P556 Applied Machine Learning
Lecture 2

D.S. Williamson



Basics of Machine Learning

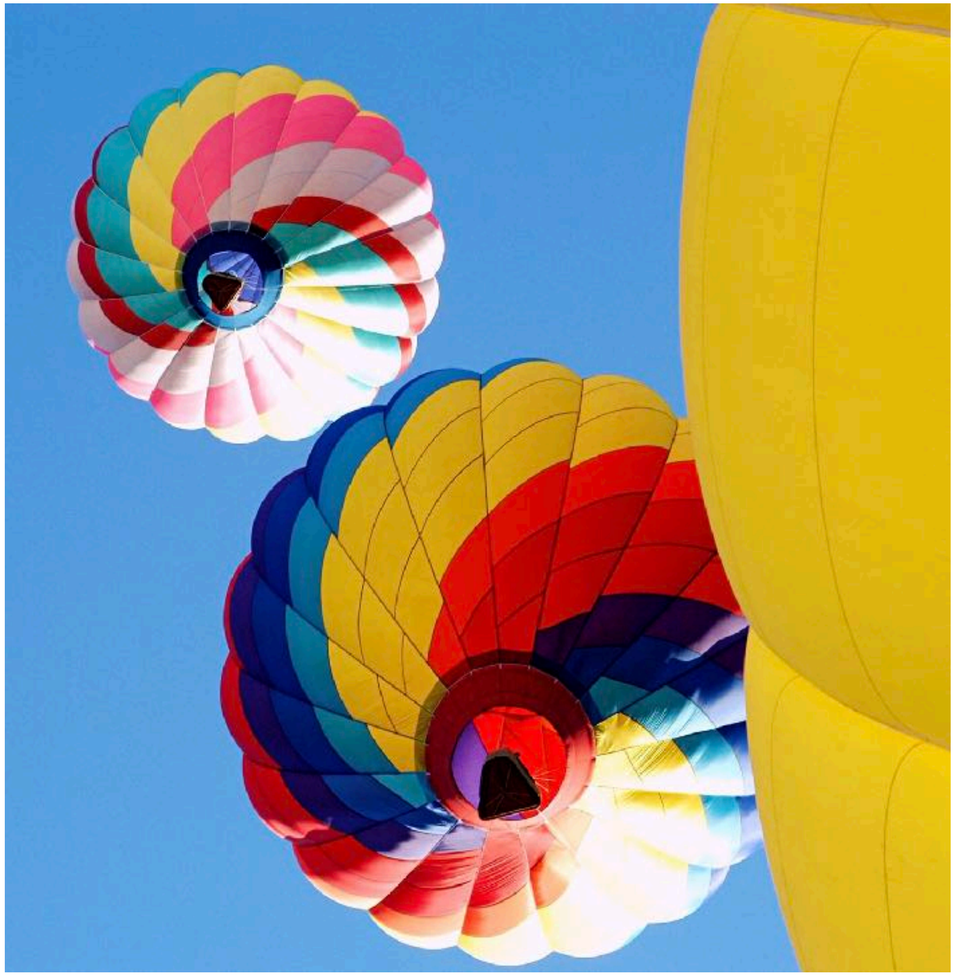
CSCI-P556 Applied Machine Learning
Lecture 2

D.S. Williamson

Agenda and Learning Outcomes

Today's Topics

- Topics:
  - Types of ML algorithms
  - Labeled Data
  - Other definitions and terms
- Learning Outcomes: At the end of today's class, students should:
  - Understand the different types of machine learning algorithms
  - Understand key definitions

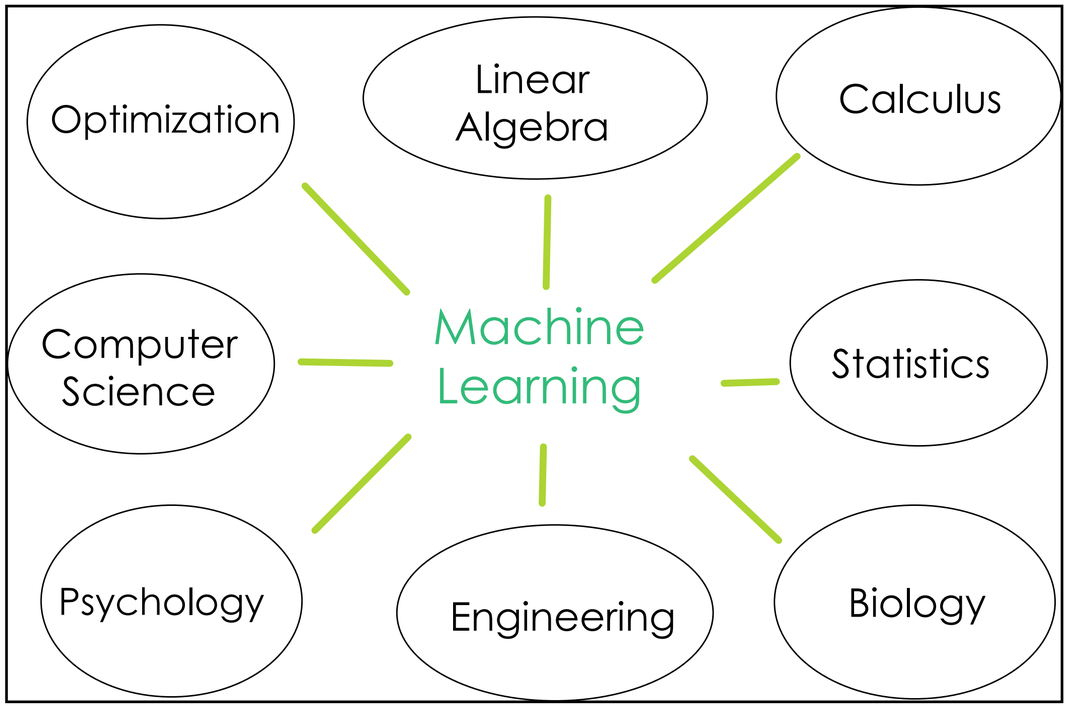


This class introduces important terms and definitions

Recall: What is Machine Learning?

It's Multifaceted

- "A set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty" ~ Kevin P. Murphy
- Concerned with developing, analyzing, and applying algorithms that make useful inferences when provided with data
- Provides a framework for solving hard problems!



Data Uncertainty

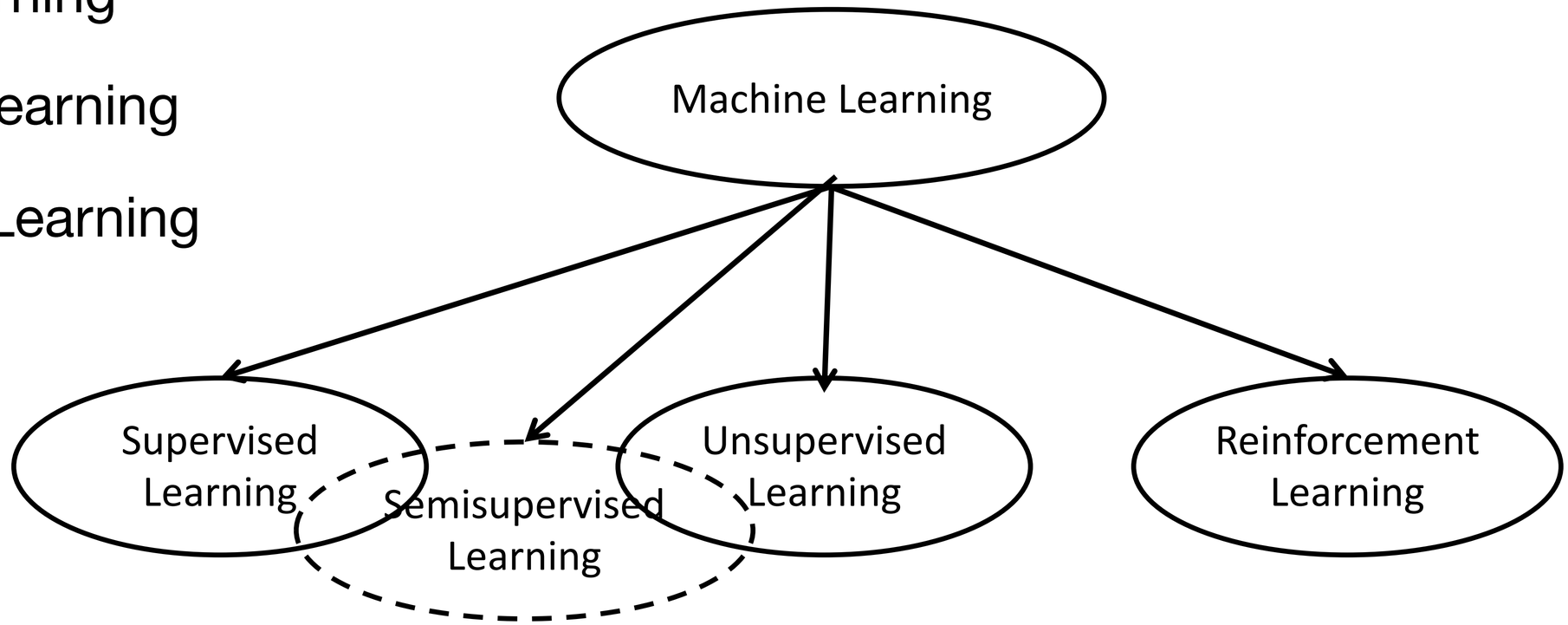
Data varies and is ambiguous

- Machine learning is needed since there is uncertainty in data
- Types of uncertainty:
  - **How can I make future predictions from past data?**
    - Weather prediction/Google searches
    - Past COVID vaccine success on future patients
  - **What is the best model or representation for this data?**
    - Probabilistic or Matrix factorization?
    - Learned embeddings or feature extraction
  - **How do I make decisions if info is missing?**
    - Packet loss/Interference

Types of Machine Learning

Three main branches of ML

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

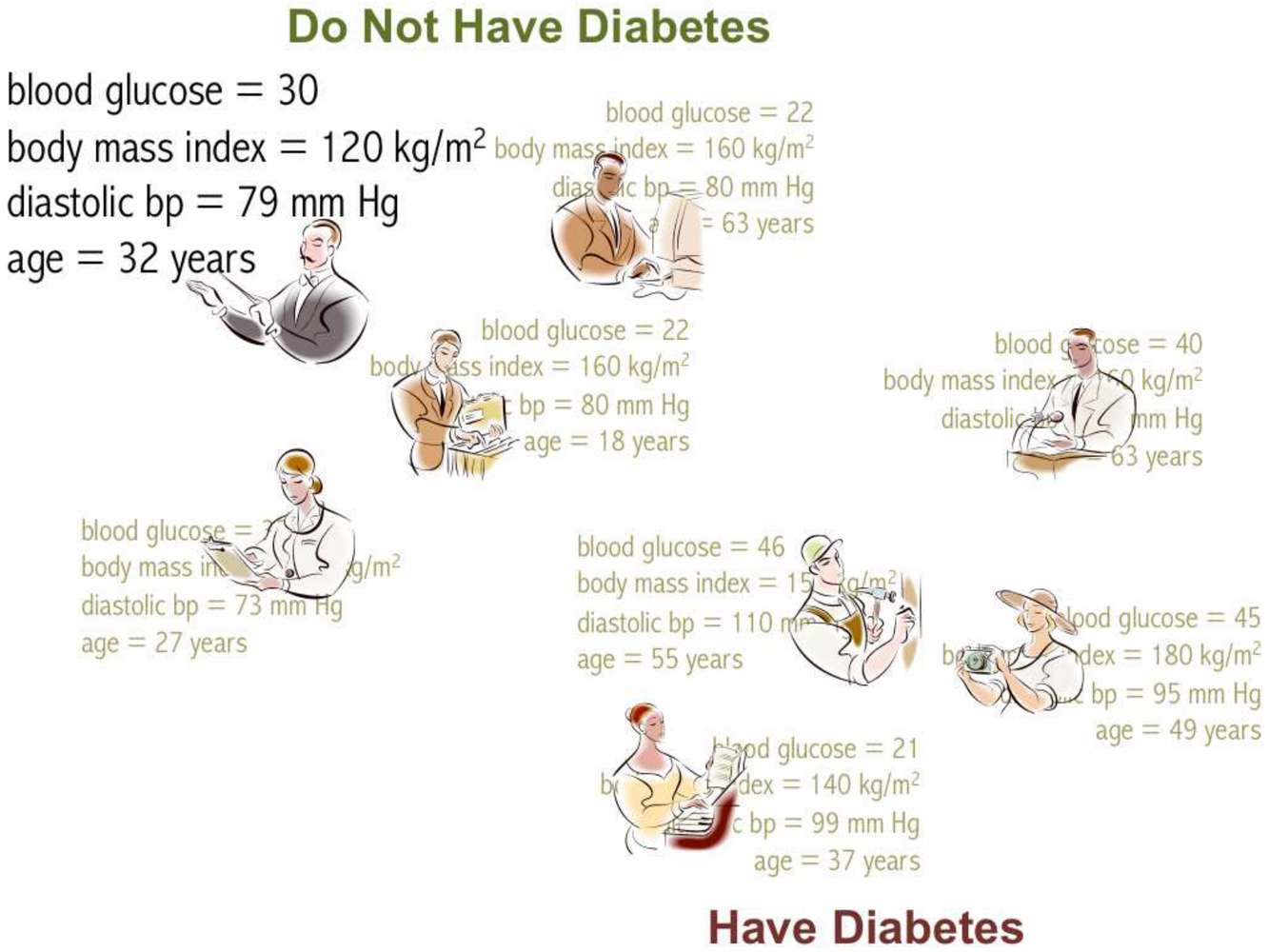


Supervised Learning - Medical Diagnosis

Probably the most researched area

- Have “supervised” or labeled information

- Often predict label information from raw data
  - No human expert is available
  - Humans can perform task but can’t describe how they do it
  - Desired function changes frequently

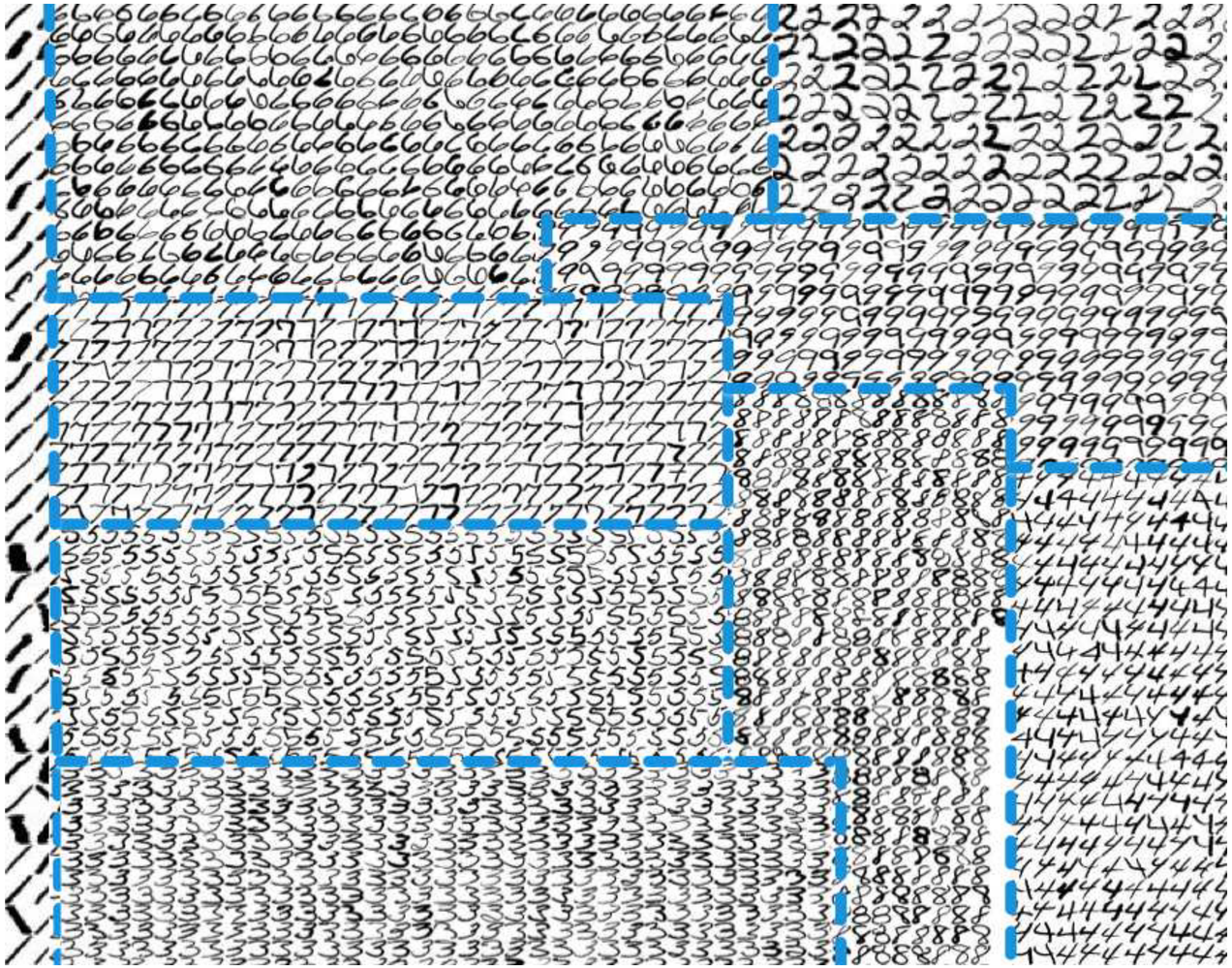


Supervised Learning - Handwritten Digit Recognition

Probably the most researched area

- Have “supervised” or labeled information

- Often predict label information from raw data
  - No human expert is available
  - Humans can perform task but can’t describe how they do it
  - Desired function changes frequently

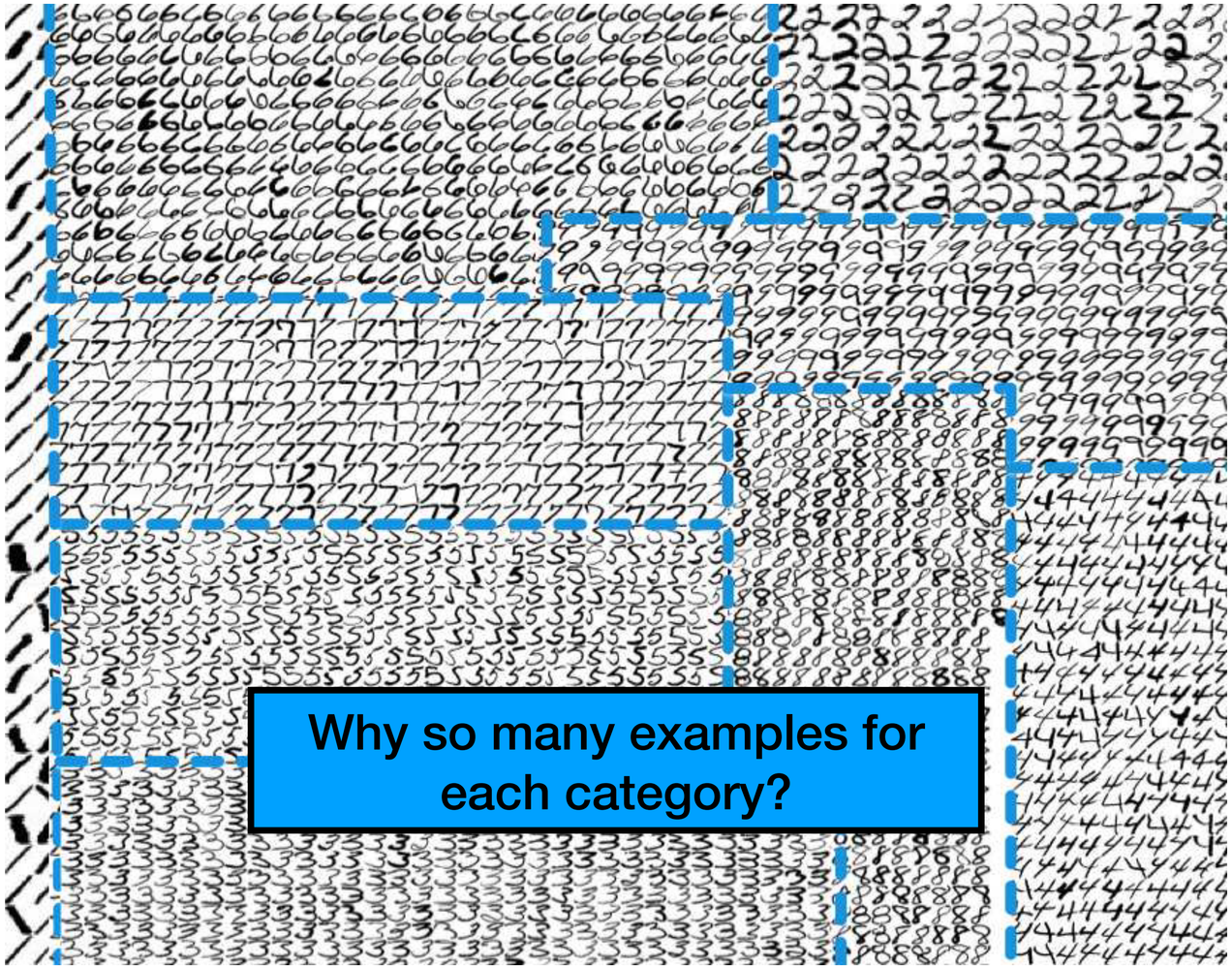


Supervised Learning - Handwritten Digit Recognition

Probably the most researched area

- Have “supervised” or labeled information

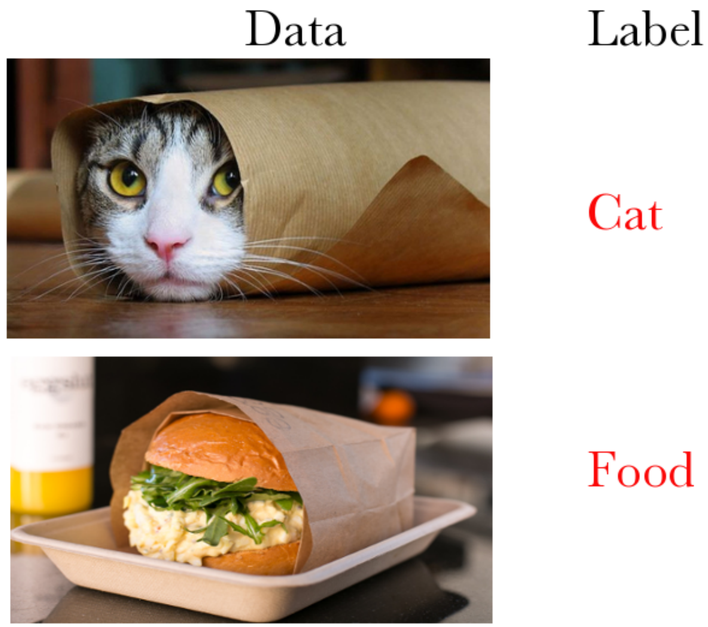
- Often predict label information from raw data
  - No human expert is available
  - Humans can perform task but can’t describe how they do it
  - Desired function changes frequently



Labeled Data for Supervised Learning

Needed for Supervised Learning

- Labeled data
  - Information about the data (image, medical record, audio signal, etc.) is known
  - Data and label combine to form a pair
  - Typically have hundreds, thousands, or millions of example pairs
  - Examples:
    - Object recognition: input image (e.g. data) and object in image (e.g. label) is known/ provided
    - Speech recognition: input audio (e.g. data) and corresponding text (e.g. label)



Labeled Data

Mathematical Notation

- Supervised learning requires labels data (e.g. observation with label)

$$D = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), (\boldsymbol{x}_2, \boldsymbol{y}_2), \ldots, (\boldsymbol{x}_i, \boldsymbol{y}_i), \ldots, (\boldsymbol{x}_N, \boldsymbol{y}_N)\}$$

x_i: i[th] observation, input, data point, feature, sample
y_i: i[th] label, target, output

- The inputs/outputs can be single (1D) or multi-dimensional
  - 1D example: a person's height or weight
  - Multi-dimensional example: an image
- **Goal**: *Learn a "functional mapping" from an input, x, to it's label, y*
  - You're given a labeled set of input-output(label) pairs, D
  - D is called the dataset
  - N is the number of pairs

Supervised Learning

- Underlying assumptions:
  - Features/inputs are easy to obtain
  - Targets/labels are difficult to observe or collect
  - A relationship exists between the features and targets

- The exact relationship or mapping is unknown! $y_i = F(x_i)$ The function *F()* is unknown
- SL uses data to learn (or approximate) this function. $y_i \approx \hat{F}(x_i)$

The function $\hat{F}()$ approximates *F()*

Labeled data for animal recognition

An Example for Supervised Learning

- **What "mapping" exists between the image and class label?**
  - What makes a cat a cat and a dog a dog?
  - They both have two eyes, two ears, and a nose.
  - But what makes them different or the same?

Cats

Dogs

Bears

Fish

Types of Labels (or Targets)

Labels are generally divided into two classes

$$D = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), (\boldsymbol{x}_2, \boldsymbol{y}_2), \ldots, (\boldsymbol{x}_i, \boldsymbol{y}_i), \ldots, (\boldsymbol{x}_N, \boldsymbol{y}_N)\}$$

- **Categorical**: Integer (Discrete) values are assigned as label
  - Examples:
    - *Fruit*: Apple, Orange, Banana, Grapes, etc.
    - *Musical artist*: Michael Jackson, Taylor Swift, Elvis, etc.
    - *Speech present*: yes or no
  - C = 2 (*Binary Classification*); C > 2 (*Multiclass classification*)
  - ***One-hot encoding*** is also done
    - Define binary vector based on number of labels
    - True label for input gets value of 1 all others get zero
      - Animal recognition problem define 4-D vector with indexing [Cats, Dogs, Bears, Fish]
      - Label vector for Dog image is [0, 1, 0, 0]

	wt [kg]	ht [m]	T [°C]	sbp [mmHg]	dbp [mmHg]	y
$\mathbf{x}_1$	91	1.85	36.6	121	75	-1
$\mathbf{x}_2$	75	1.80	37.4	128	85	$+1$
$\mathbf{x}_3$	54	1.56	36.6	110	62	-1

Table 3.1: An example of a binary classification problem: prediction of a disease state for a patient. Here, features indicate weight (wt), height (ht), temperature (T), systolic blood pressure (sbp), and diastolic blood pressure (dbp). The class labels indicate presence of a particular disease, e.g. diabetes. This data set contains one positive data point ($\mathbf{x}_2$) and two negative data points ($\mathbf{x}_1$, $\mathbf{x}_3$). The class label shows a disease state, i.e. $y_i = +1$ indicates the presence while $y_i = -1$ indicates absence of disease.

$$y \in \{1, \ldots, C\}$$

Types of Labels (or Targets)

Labels are generally divided into two classes

$$D = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), (\boldsymbol{x}_2, \boldsymbol{y}_2), \ldots, (\boldsymbol{x}_i, \boldsymbol{y}_i), \ldots, (\boldsymbol{x}_N, \boldsymbol{y}_N)\}$$

- **Regression**: Decimal (Continuous) values are assigned as the label
  - Examples:
    - A person's height or weight to the 3rd decimal place
    - The cost of a home
    - Stock market price
    - Outputting an image of a dog/cat/bear/fish
    - Create musical audio signals

	size [sqft]	age [yr]	dist [mi]	inc [$]	dens [ppl/mi^2]	y
$\mathbf{x}_1$	1250	5	2.85	56,650	12.5	2.35
$\mathbf{x}_2$	3200	9	8.21	245,800	3.1	3.95
$\mathbf{x}_3$	825	12	0.34	61,050	112.5	5.10

Table 3.2: An example of a regression problem: prediction of the price of a house in a particular region. Here, features indicate the size of the house (size) in square feet, the age of the house (age) in years, the distance from the city center (dist) in miles, the average income in a one square mile radius (inc), and the population density in the same area (dens). The target indicates the price a house is sold at, e.g. in hundreds of thousands of dollars.

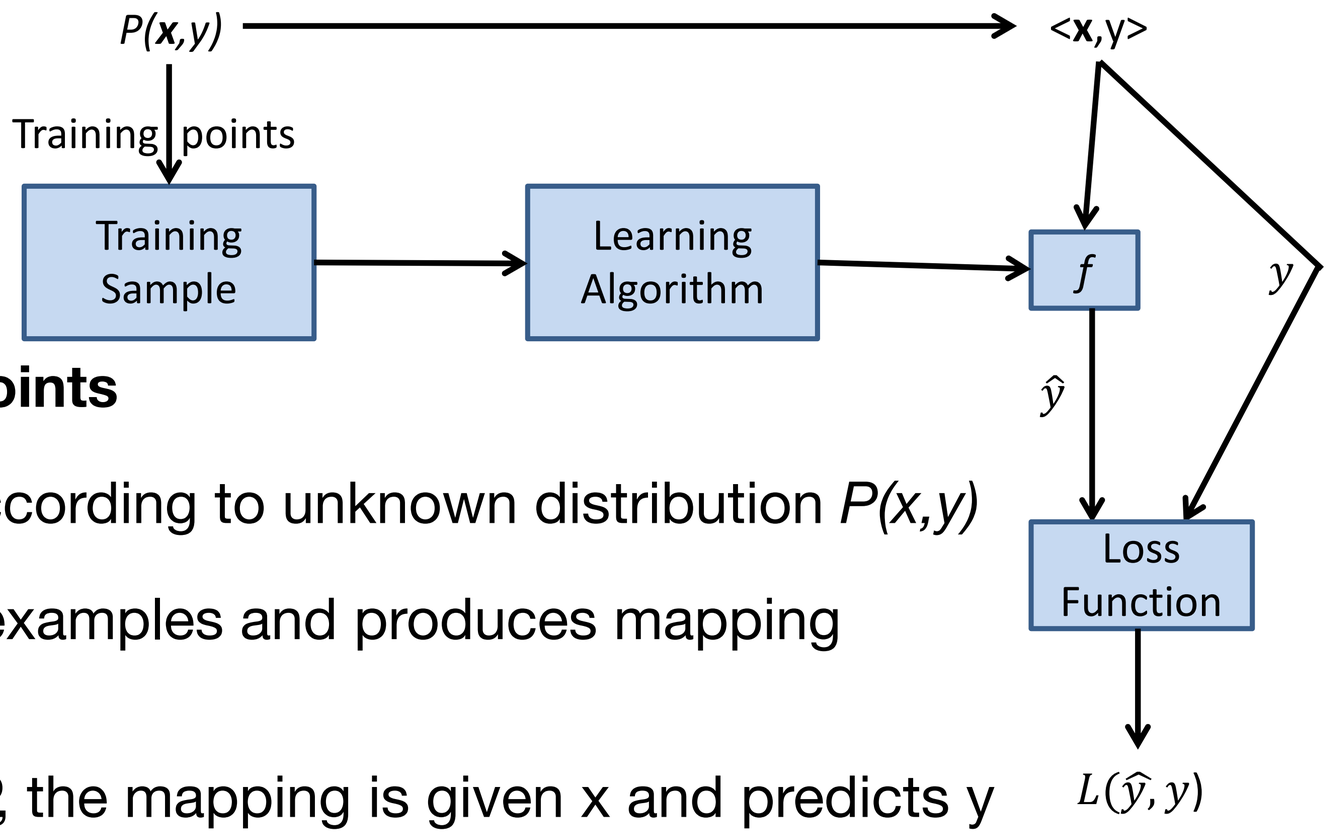
- It is termed ***regression*** when a supervised learning algorithm learns a mapping from an input to a continuous label

$y \in \mathbb{R}$

Real number (e.g. decimal or float; 1.232,343,232.4545,…)

Supervised Learning - The Learning Process

Approximating the function



- **Divide data into training and testing points**
  - *Training examples*: randomly drawn according to unknown distribution *P(x,y)*
  - *Learning algorithm*: analyzes training examples and produces mapping function f

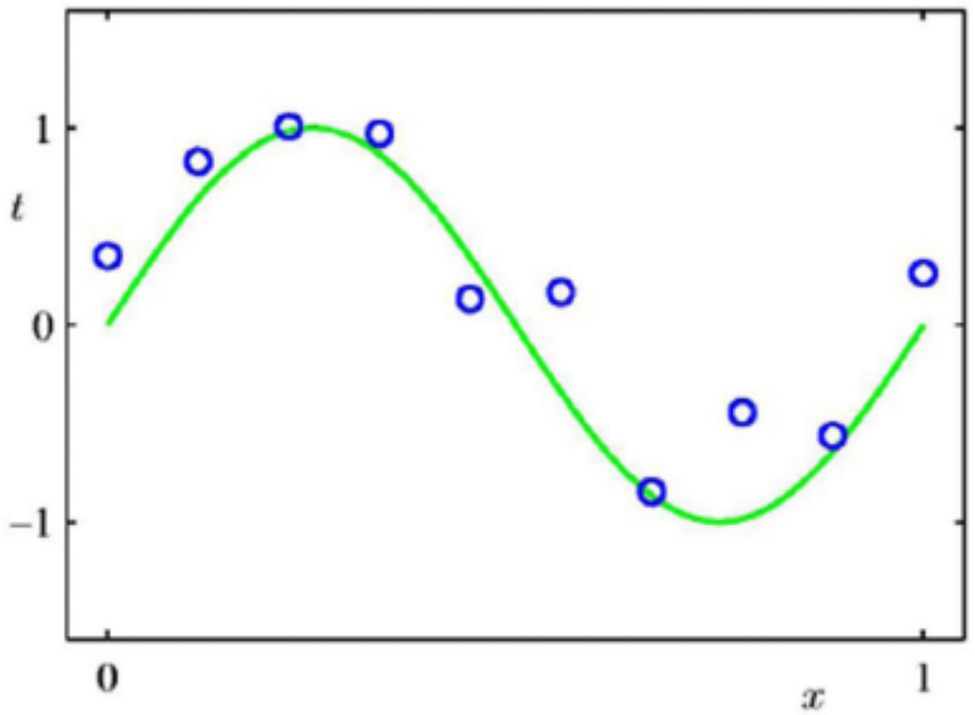
$$\hat{y} = f(x)$$

  - Given a new point <x,y> drawn from P, the mapping is given x and predicts y
  - The loss (or error) is measured by L(ŷ,y)
  - **Goal**: Find the f that minimizes the expected loss: E[L(f(x),y)]

Example - Curve fitting

Supervised Learning

- Goals:
  - Find the “best” curve to fit the data (in blue)
  - Make ***accurate*** predictions on new **unseen** points such that some notion of error is minimized
  - Predict values of *t* given values of *x*
- Input:
  - The Blue circles
  - Noise makes this more difficult



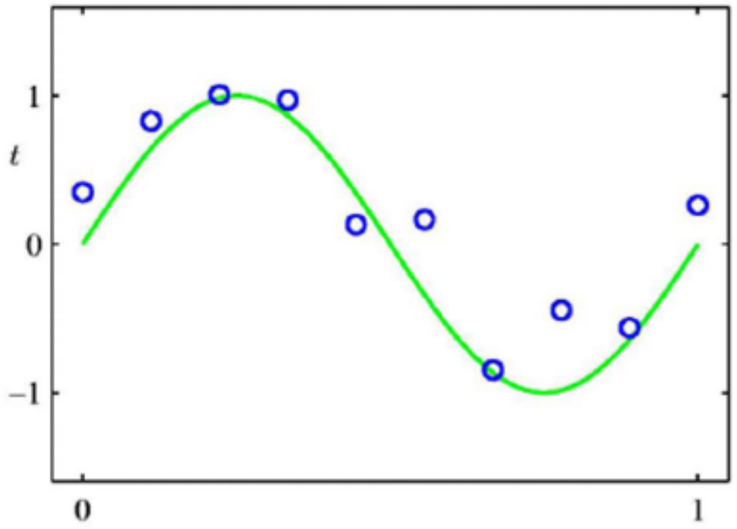
The underline function:

$$t = \sin(2\pi x) + \varepsilon$$

where ε is Gaussian noise

Example - Curve fitting

Supervised Learning



The underline function:

$t = \sin(2\pi x) + \varepsilon$

where ε is Gaussian noise

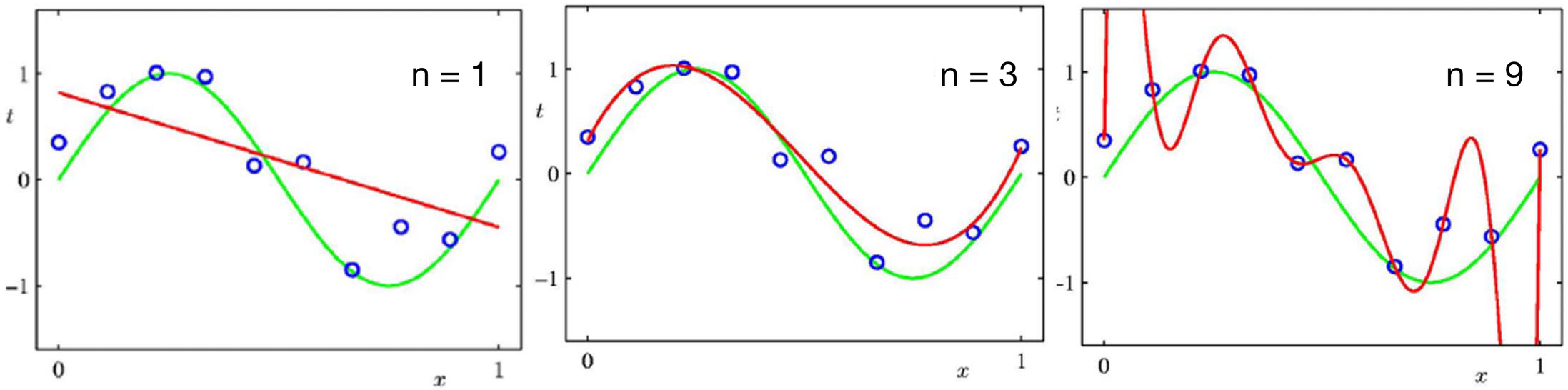
- There are ***infinitely*** many possible functions that can fit this data.
- Hence, we need to focus on a smaller number of functions possible
  - This is called ***hypothesis*** space
- Ex., we can restrict the order of polynomial (nth order)

$$f(x, w) = w_0 + w_1 x + w_2 x^2 + \ldots + w_n x^n$$

  - We wish to learn the parameters $\langle w_0, w_1, \ldots, w_n \rangle$
- These parameters must be learned such that some loss function is minimized
  - For example, squared error $E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}(y(x_n, \mathbf{w}) - t_n)^2$
  - Let us not worry about how it is solved yet. These can be solved easily by optimization techniques.

Which Model to choose?

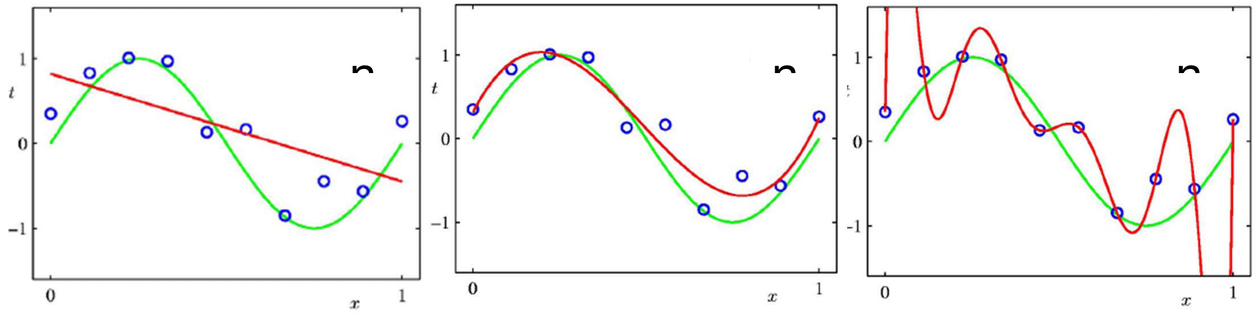
Different Supervised Learning Models



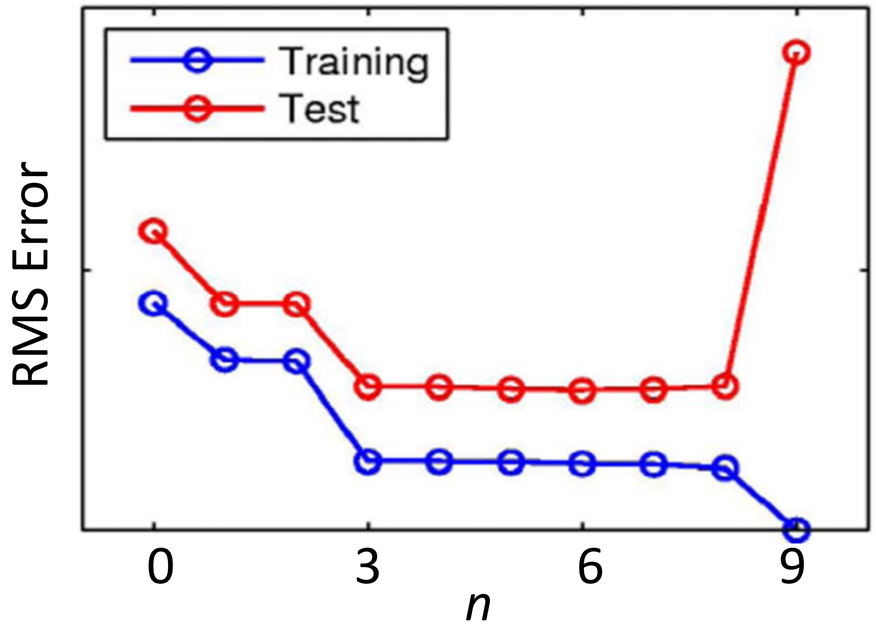
- Red curve is the one with different values for n
- Which *n* should we choose? – Model selection
- Can we use the error E(w) as the criterion?

Overfitting

Problem with Supervised Learning



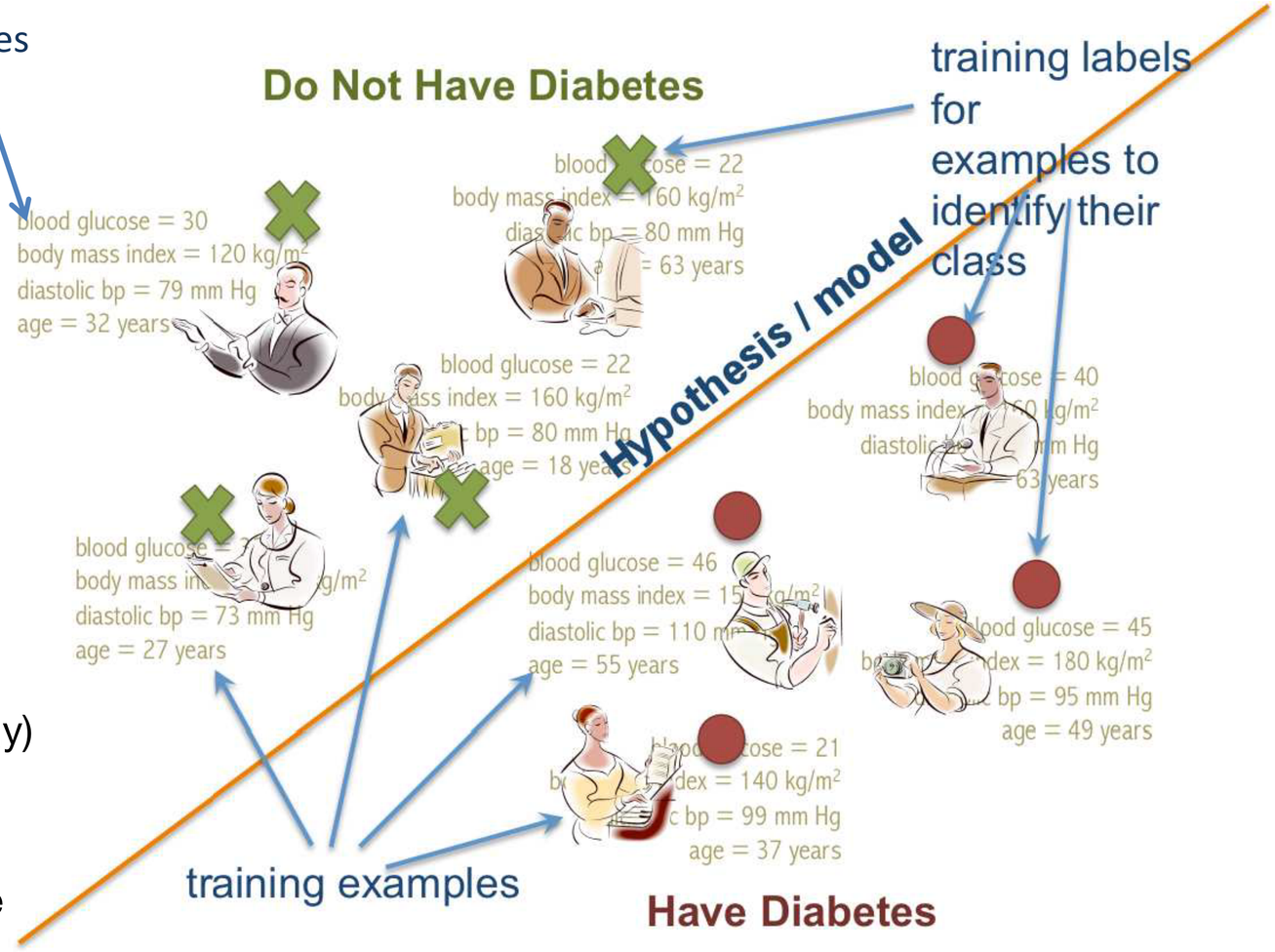
- As *n* increases, training error decreases monotonically
- As *n* increases test error can increase
- Test error can decrease at first, but increases
- ***Overfitting*** can occur:
  - ***When the model is too complex and trivially fits the data (i.e., too many parameters)***
  - When the data is not enough to estimate the parameters
  - **Model captures the noise (or the chance):**
    - Unintended correlations between input and label
    - Correlations specific to the training data



Terminology

Supervised Learning

- ***Training Example***: <x,y>
  - x: feature vector (input data)
  - y: label
    - Continuous valued for regression
    - [1,2,…,C] for classification
- ***Training set***
  - A set of training examples drawn randomly from P(x,y)
  - ***Key assumption***: Independent and identically distributed. i.e., all the examples are drawn from the same distribution, but are drawn independent of one another
- ***Target function***: True (unknown) mapping form x to y



Terminology

Supervised Learning

- ***Hypothesis***: A function h considered by the learning algorithm to be similar to the target function
  - *Regression*: could be prediction model
  - *Classification*: boundary line(s)

- ***Test set***: A set of examples drawn from P(x,y) to evaluate the “goodness of h”

- ***Hypothesis Space***: The space of all hypotheses that can in principle be considered and returned by the learning algorithm

Key function approximation approaches

Supervised Learning

- **Directly learn a mapping y = f(x)**
  - No uncertainty is captured
- **Learn the joint distribution i.e., learn p(y,x)**
  - Captures uncertainty about both the attributes **x** and the target y
- **Learn the conditional distribution i.e., learn p(y|x)**
  - p(**x**,y) = p(y|**x**)p(**x**)
  - Hence this avoids modeling the distribution of **x**
  - In general, this is akin to assuming an uniform distribution over **x**
  - Can also be considered as saying “I do not care about **x** but only P(y|**x**)
- **Once we learn p, how do we choose y? This is called as decision-theory**

Key Questions

Supervised Learning

- What are good hypothesis spaces?
- What algorithms work on these spaces?
- How can we generalize to unseen points (i.e., avoid overfitting)?
- How can we trust our results?
- Are some problems computationally intractable?
- How can we formulate practical problems as Machine Learning ones?

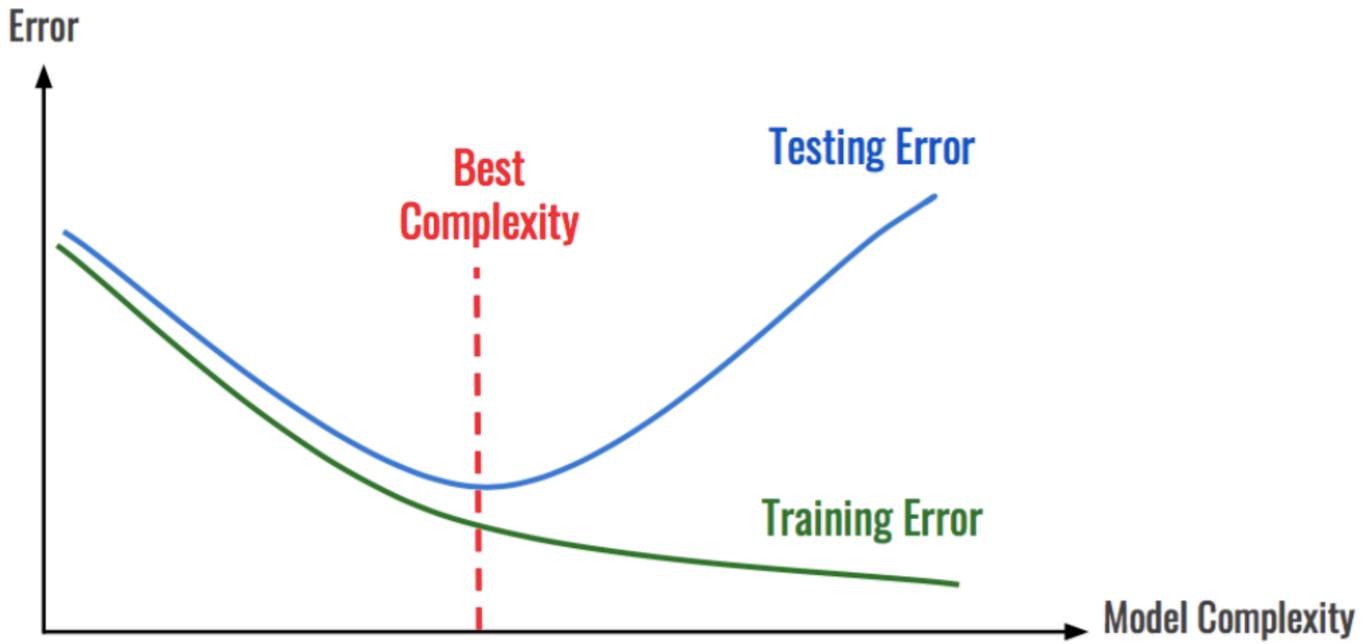
Learning Algorithms

Different Strategies

- **Search**
  - Direct Computation
  - Local Search: Start with an initial hypothesis, make gradual improvements til a local maximum is reached
  - Constructive search: start with an empty hypothesis and grow it till local maximum
- **Timing**
  - Eager: Analyze training data and construct hypothesis
  - Lazy: Store the training data and wait til a test point is presented, then construct an hypothesis to classify that test point
- **Online vs. Batch (for eager)**
  - Online: Analyze each training point as they arrive
  - Batch: Collect all examples, analyze them together
- **Parametric vs. Non-Parametric Models**
  - Parametric: model has a fixed number of parameters
  - Non-Parametric: number of parameters grow with the amount of training data

Model Selection

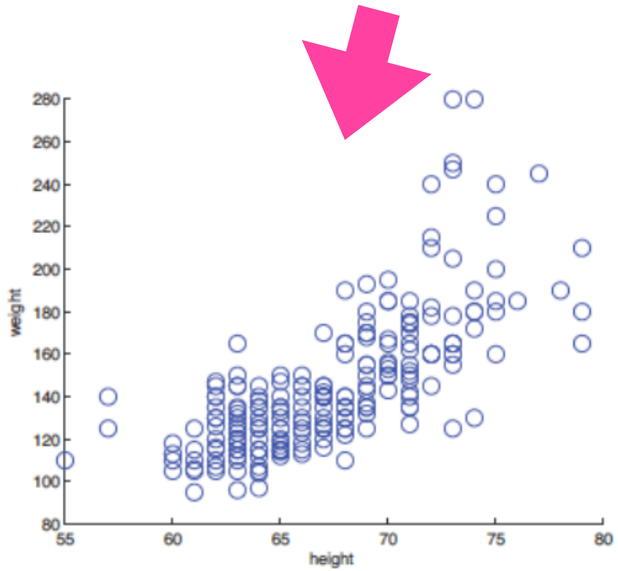
Which version of the model do I use?



Unsupervised Learning

Discover patterns or structure in the data

Input Data



(a)

- Only have (or use) the data information (e.g. ignore labels)

$$D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_i, \ldots, \boldsymbol{x}_N\}$$

- Examples for unsupervised learning
  - **Clustering**: K-means, vector quantization, Gaussian mixture models
  - **Dimensionality Reduction**: principal components analysis, nonnegative matrix factorization
  - **Topic Modeling:** often used in NLP

Unsupervised Learning

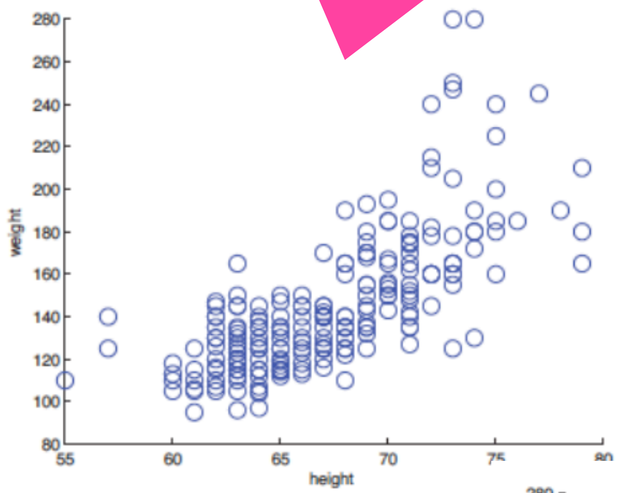
Discover patterns or structure in the data

- Only have (or use) the data information (e.g. ignore labels)

$$D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_i, \ldots, \boldsymbol{x}_N\}$$

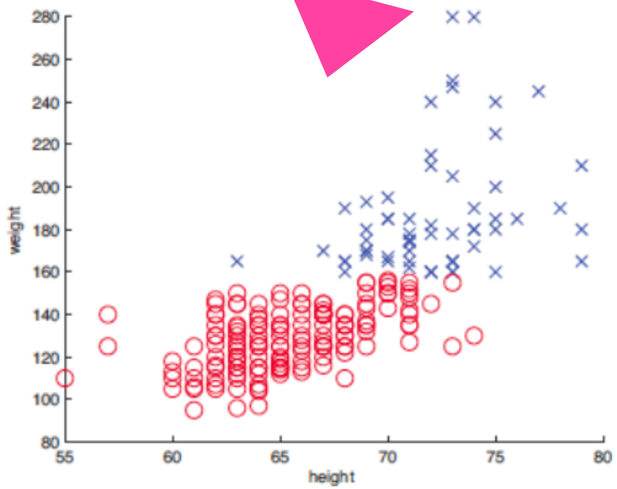
- Examples for unsupervised learning
  - **Clustering**: K-means, vector quantization, Gaussian mixture models
  - **Dimensionality Reduction**: principal components analysis, nonnegative matrix factorization
  - **Topic Modeling:** often used in NLP

Input Data



(a)

Date Clustered into two Groups



(b)

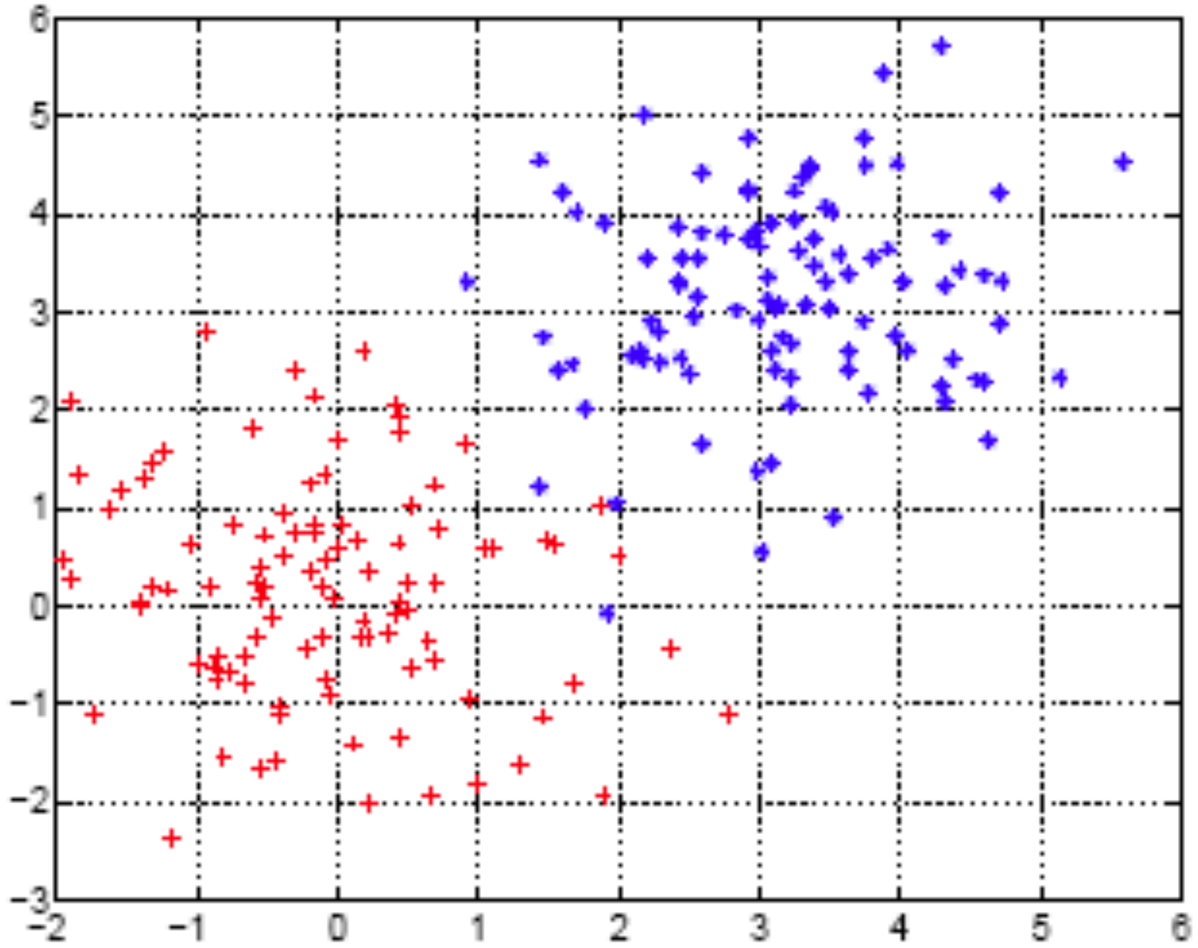
Unsupervised Learning

Discover patterns or structure in the data

- Only have (or use) the data information (e.g. ignore labels)

$$D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_i, \ldots, \boldsymbol{x}_N\}$$

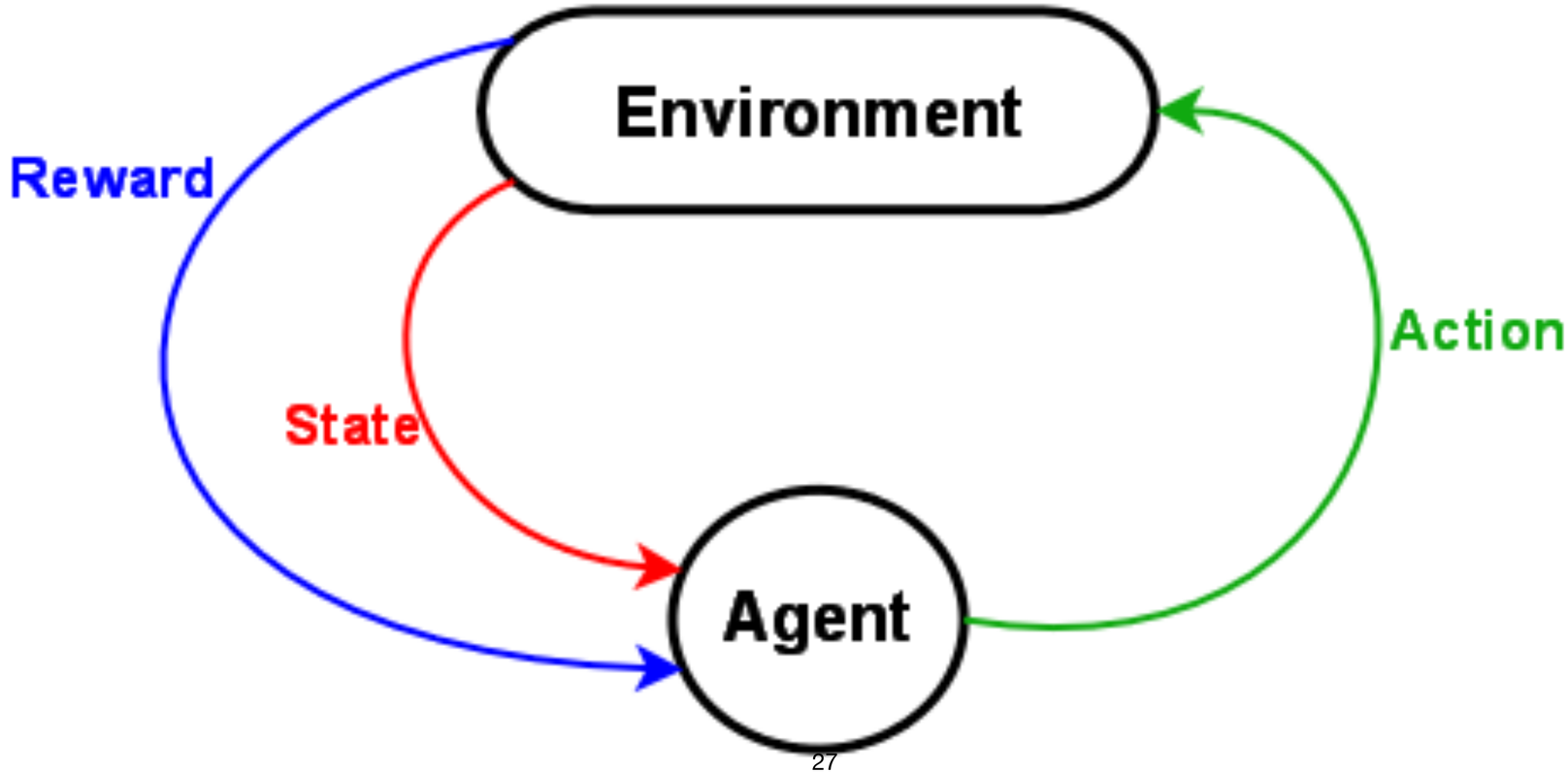
- Arguably more in line with human and animal learning
  - No one “teaches” a baby how to walk
  - No one “teaches” an animal how to run



Reinforcement Learning

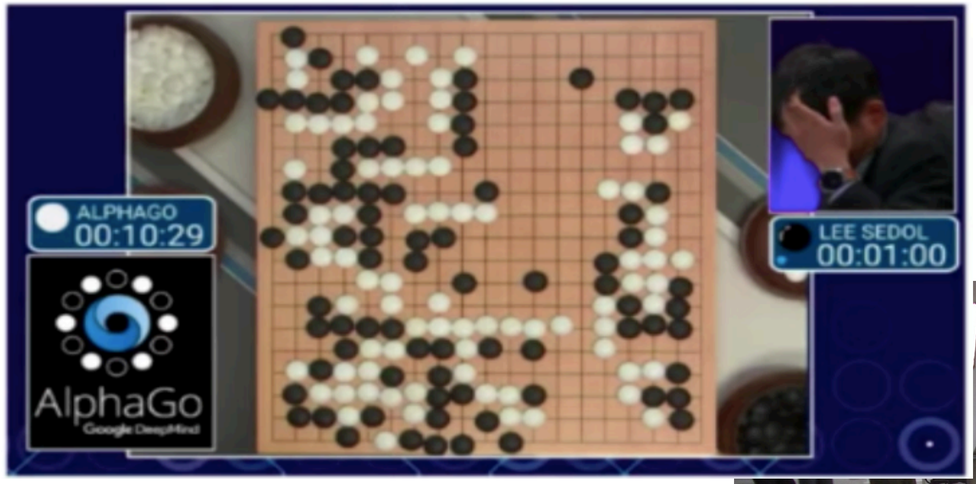
Learn from environment and rewards

- Agent leans how to “act” when given rewards (or punishments)



Reinforcement Learning

Examples



Types of Machine Learning

Other Types exist

- Semi-supervised
- Active Learning
- Forecasting
- …

Knowing Your Goal and Your Data

Other Considerations

- What question(s) am I trying to answer? Do I think the data collected can answer that question?

- What is the best way to phrase my question(s)?

- Have I collected enough data to represent the problem I want to solve?
  - Plotting Your Data !!!

Knowing Your Goal and Your Data

Other Considerations

- What features of the data did I extract, and will these enable the right predictions?

- How can I measure success in my application?

- Can I interpret the model and the process to someone else?

Sidebar: Ethical Considerations

- Privacy
- Fairness, bias, ethics
- Benefit vs. Harm
- …

Next Week

Data Pre-Processing

HW0 - Environment Setup

Read