Where we are

- Informatics
  - management, manipulation, integration
  - emphasis on scale, some emphasis on tools
- Analytics
  - statistical estimation and prediction
  - machine learning, data mining
- Visualization
  - communication and presentation

What is Machine Learning?

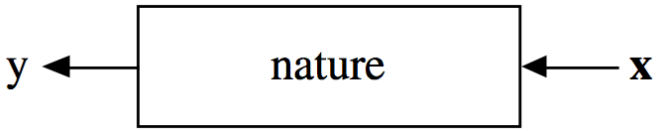
- "Systems that automatically learn programs from data" [Domingos 2012]

- Teaching a computer about the world

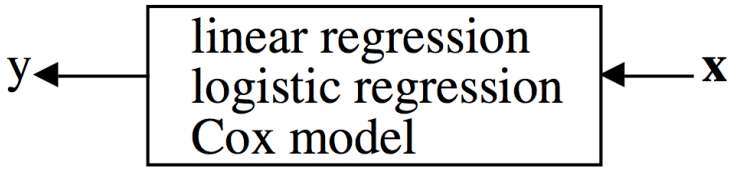
[Mark Dredze]

What's the difference between Statistics and Machine Learning?

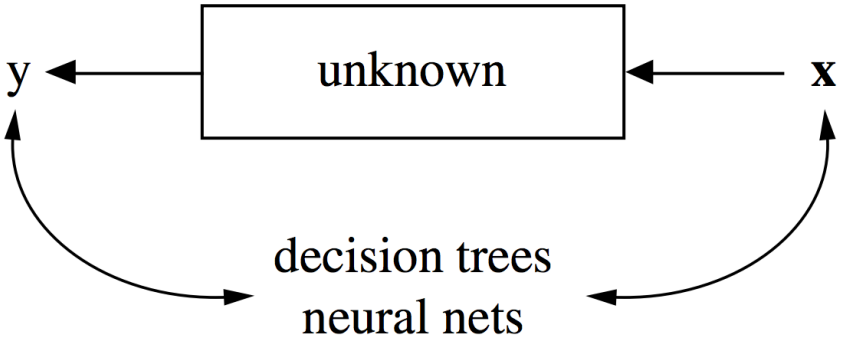
One view:



Emphasis on stochastic models of nature:



Find a function that predicts y from x:
no model of nature implied or needed



Leo Breiman, Statistical Modeling: The Two Cultures, Statistical Science 16(3), 2001

[Witten]

Toy Example

Goal: Predict when we play

outlook	temperature	humidity	windy	PLAY?
sunny	hot	high	false	no
sunny	hot	high	true	no
overca				yes
rainy				yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny				no
sunny				yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

hypothesis: we only play when its sunny?

No

hypothesis: we don't play if its rainy and windy?

No

Terminology

- **classification**
  - The learned attribute is categorical (“nominal”)
- **regression**
  - The learned attribute is numeric

Terminology

- Supervised Learning ("Training")
  - We are given examples of inputs and associated outputs
  - We learn the relationship between them
- Unsupervised Learning (sometimes: "Mining")
  - We are given inputs, but no outputs
    - unlabeled data
  - Learn the "latent" labels
  - Ex: Clustering, dimension reduction

Example: Document Classification

"The Falcons trounced the Saints on Sunday"

Sports

"The Mars Rover discovered organic molecules on Sunday"

Science

How do we set this up?
What are the rows and columns of our decision table?