

# INTRO to DATA SCIENCE

## LECTURE 2: MACHINE LEARNING

- 1.WHAT IS MACHINE LEARNING?**
- 2.MACHINE LEARNING ALGORITHMS**
- 3.DATAEXPLOR01 REVIEW**
- 4.PYTHON TOOLS FOR ML**
- 5.LAB: PRACTICE**

# I. WHAT IS MACHINE LEARNING?

*“Machine learning, a branch of artificial intelligence, is about the construction and study of systems that can learn from data.”* (source: [http://en.wikipedia.org/wiki/Machine\\_learning](http://en.wikipedia.org/wiki/Machine_learning))

*"Machine learning, a branch of artificial intelligence, is about the construction and study of systems that can learn from data."* (source: [http://en.wikipedia.org/wiki/Machine\\_learning](http://en.wikipedia.org/wiki/Machine_learning))

- ▶ *"Field of study that gives computers the ability to learn without being explicitly programmed"*

Arthur Samuel

Machine Learning Pioneer

Samuel checkers playing program

*“Machine learning, a branch of artificial intelligence, is about the construction and study of systems that can learn from data.”* (source: [http://en.wikipedia.org/wiki/Machine\\_learning](http://en.wikipedia.org/wiki/Machine_learning))

- ▶ *“Field of study that gives computers the ability to learn without being explicitly programmed”*

Arthur Samuel

Machine Learning Pioneer

Samuel checkers playing program

- ▶ *“The automatic discovery of regularities in data through the use of computer algorithms, and with the use of these regularities to take actions such as classifying the data into different categories”*

Christopher Bishop

Distinguished Scientist Microsoft Research

Head of Machine Learning, Cambridge, UK

---

## QUESTION

---

WHAT IS  
MACHINE LEARNING  
USED FOR?

**Prediction**

**Pattern Recognition**

**Diagnostics**

**Search Engines**

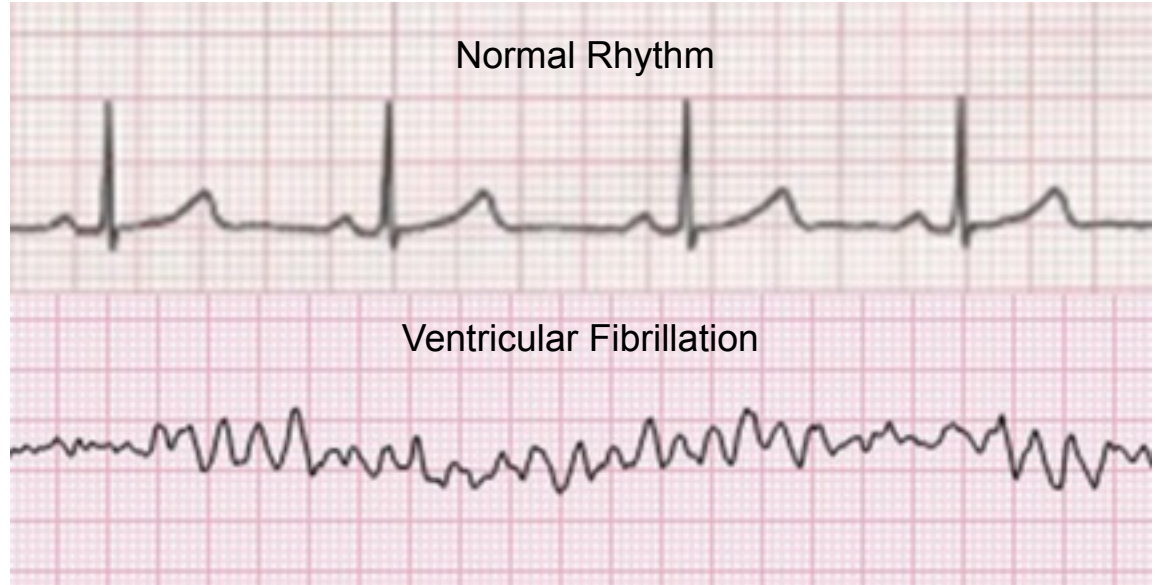
**Bioinformatics**

**Summarization**

**Machine Translation**



Automatic External Defibrillator



Johns Hopkins University School of Medicine estimates 522 lives are saved annually

- ▶ Finding patterns in (large) data sets
- ▶ Scaling out human decision making

- ▶ Algorithms vary in their ability to generalize over patterns
- ▶ Possibility of under/over-generalizing
- ▶ Limited by available data

# II. MACHINE LEARNING ALGORITHMS

1. Supervised

2. Unsupervised

# 1. Supervised

➤ *Making predictions*

Classification

Regression

# 2. Unsupervised

# 1. Supervised

➤ *Making predictions*

Classification

Regression

# 2. Unsupervised

➤ *Extracting Structure*

- Outcome measurements:  $y$  – “***targets or outputs***”
  - (also called the dependent variable, the response)
- Set of training data:  $x$  – “***features or inputs***”
  - (also called regressors, covariates, independent variables, predictor measurements)
- In ***regression***,  $y$  has real values - (54.9, 37.2, 24.6, ...)
  - (e.g. house price, temperature)
- In ***classification***,  $y$  has finite values - (0, 1)
  - (e.g. survived/died, normal rhythm/fibrillation, cat/dog/horse)



On the basis of training data

$(x_1, y_1)$

$(x_2, y_2)$

$(x_3, y_3)$

$\vdots$

$(x_N, y_N)$

we would like to:

- accurately predict unseen test cases
- understand which inputs affect the outcome, and how
- assess the quality of our predictions and inferences

- no target variable - just a set of features derived from the training data
- objective is less clear - find groups of features that behave similarly
- difficult to know how well you are doing
- can be useful as a pre-processing step for supervised learning

- spam filtering

- spam filtering
- character recognition

- spam filtering
- character recognition
- document clustering

- spam filtering
- character recognition
- document clustering
- fraud detection

- spam filtering
- character recognition
- document clustering
- fraud detection
- deciphering animal “speech”!

- III. DATAEXPLOR01 REVIEW
- IV. PYTHON TOOLS FOR ML
- V. LAB: PRACTICE