Lecture 8: Machine Learning Models

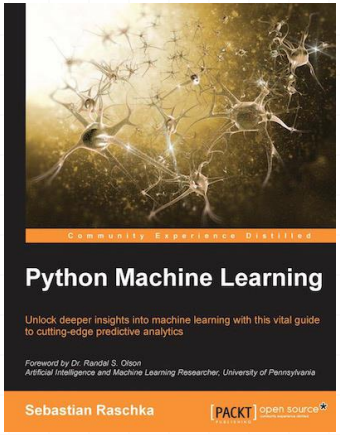
Course: Biomedical Data Science

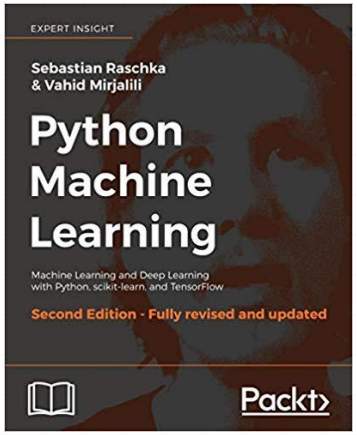
Parisa Rashidi

Fall 2018

Disclaimer

The following slides are partially based on:



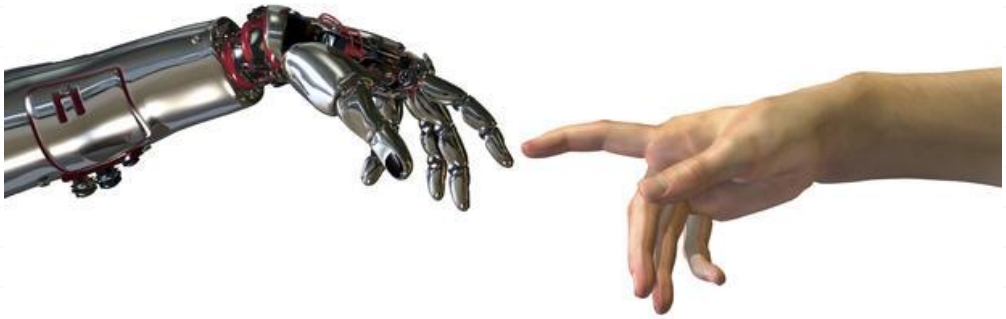


Agenda

- Machine learning inroduction
- Simple classification models
  - KNN, decision trees
- More advanced models
  - XGBoost
  - SVM
- Deep learning

Artificial Intelligence

- Artificial Intelligence (AI) has many subfields
  - Machine Learning (ML)
  - Natural Language Processing (NLP)
  - Vision
  - Robotics
  - …



What is "Learning" ?

- Machine learning is programming computers to optimize a performance criterion in a certain task using example data (i.e. past experience).

- Example task: predicting if there will be any complication 30 days after surgery
  - Performance Criteria: Number of cases correctly predicted
  - Example data: patients' medical history + outcome after 30 days

Capturing Informal Knowledge: Early Days

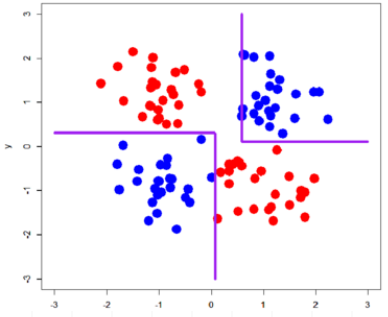
- We need to get informal knowledge to computers
  - Several systems tried to hard-code this knowledge
    - Knowledge-base approach
    - Example: Cyc, the world's longest-lived artificial intelligence project
      - A knowledgebase of the basic concepts and "rules of thumb" about how the world works

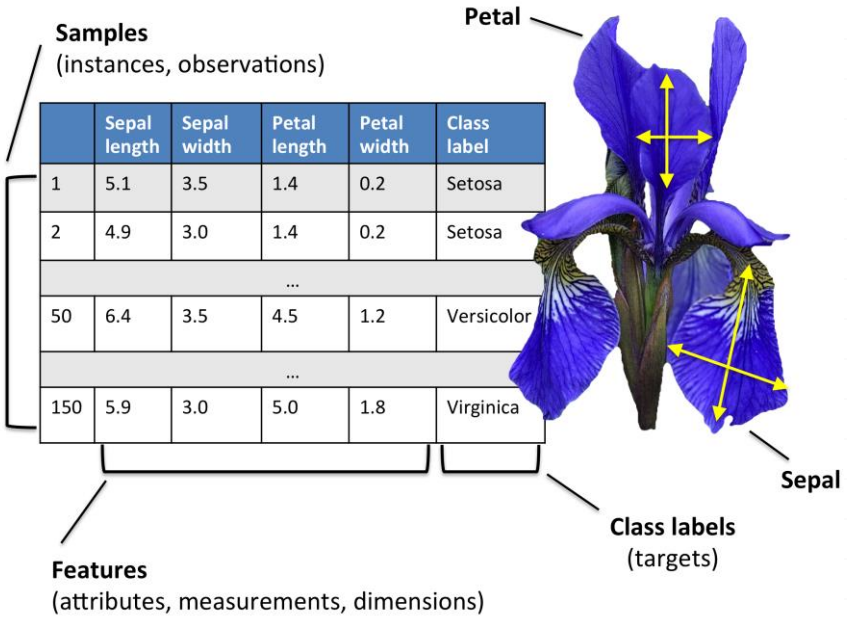
Capturing Informal Knowledge: Modern Approach

- Machine learning
  - Instead of dictating rules, let's provide data to the machine and let it learn from data.
  - Even simple algorithms might work: deciding if C-section is needed using logistic regression (Mor-Yosef et al., 1990)

Key Terms

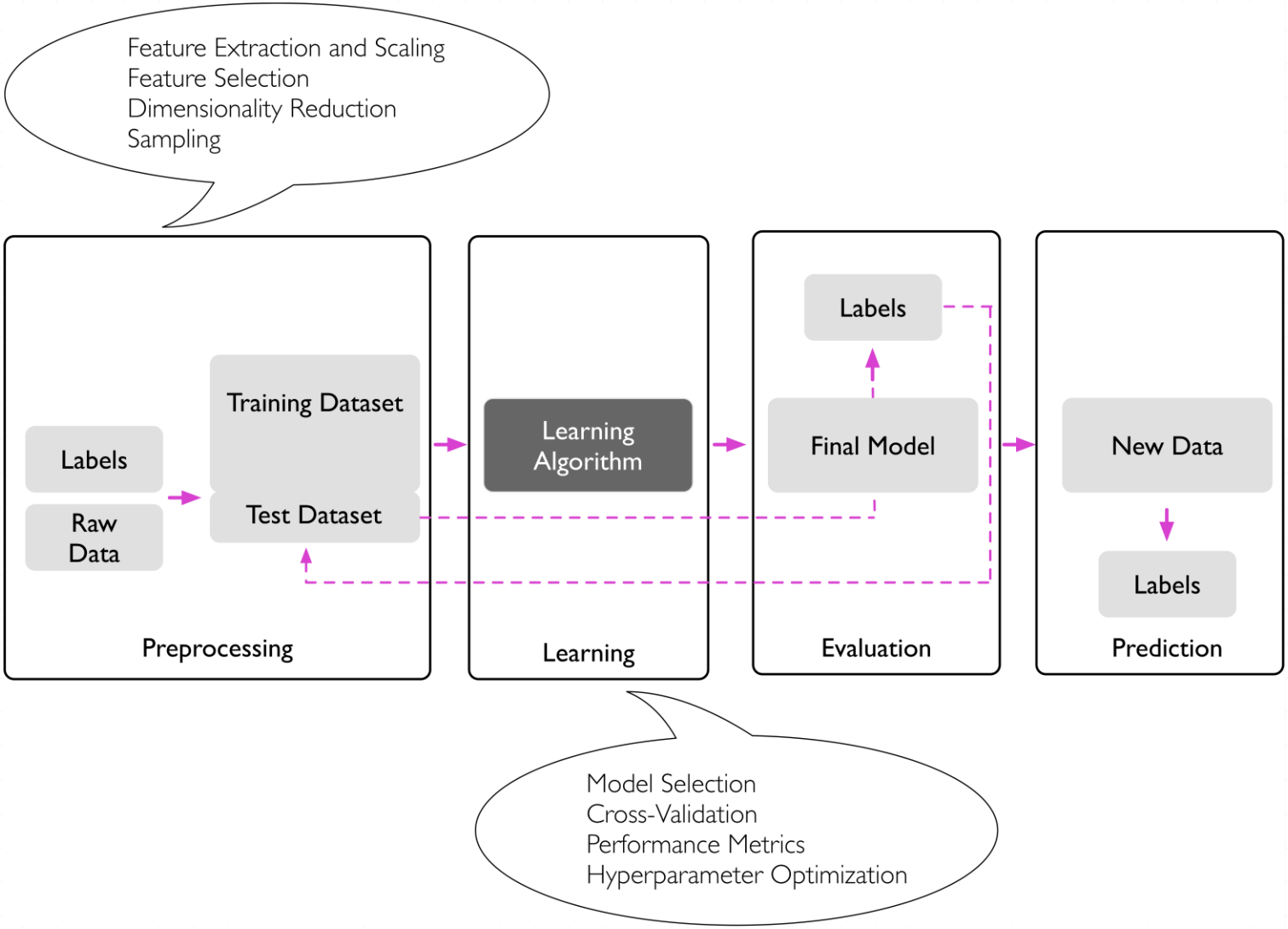
- Instance = example = data point
- Feature = independent variable
- Class label = dependent variable
- Decision boundary = separates examples in different classes





	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

Roadmap

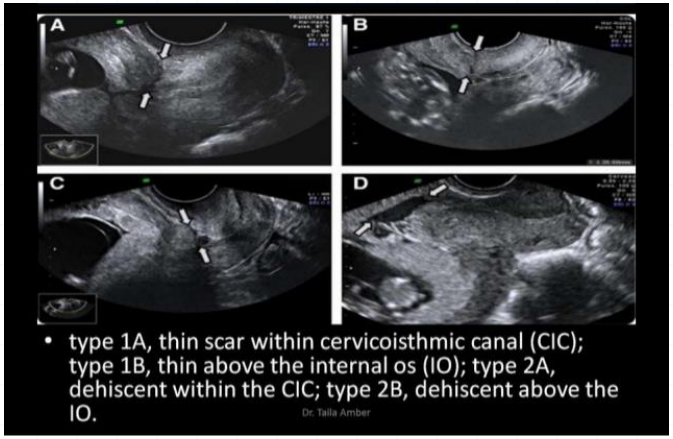

Feature Extraction and Scaling
Feature Selection
Dimensionality Reduction
Sampling
Labels
Raw Data
Training Dataset
Test Dataset
Preprocessing
Learning Algorithm
Learning
Labels
Final Model
Evaluation
New Data
Labels
Prediction
Model Selection
Cross-Validation
Performance Metrics
Hyperparameter Optimization

Representation

- Each piece of information included in the representation of the patient is known as a feature.
- The algorithm will learn how the features are correlated with the outcome.

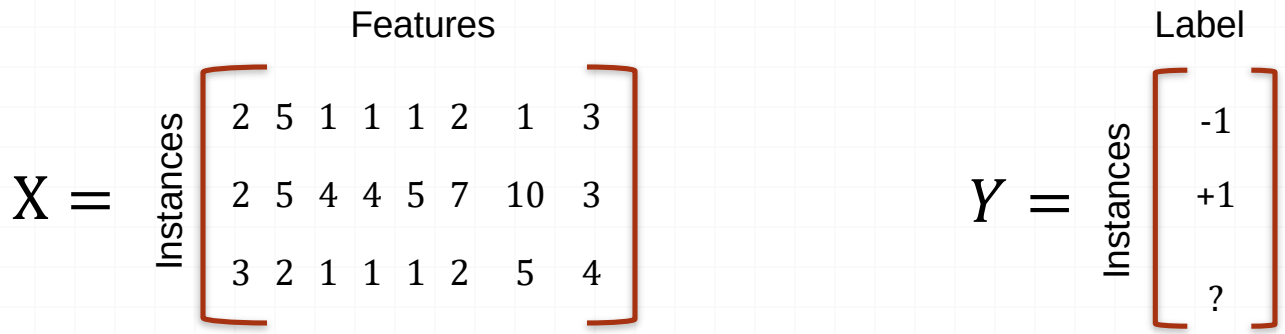
	age	Previous pregnancies	Scar	C-section
P1	21	0	0	0
P2	39	2	1	1
P3	36	1	1	?

or



Data Representation

- We usually represent data in a matrix



Note: We can also assign a probability to each label (we'll discuss it later)

Input Representation

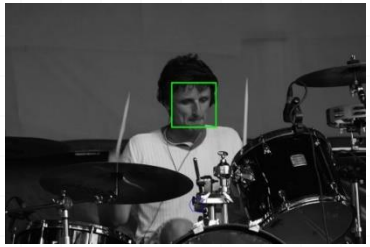
- The most common type
  - Simple records in Tables
  - Can be analyzed using regular machine learning techniques.
  - Most other data types are converted to this type (not always, we will later talk about deep learning).

ID	WGT	HGT	Cholesterol	**Risk (Class)**
1	high	short	260	high
2	high	med	254	high
3	high	tall	142	med

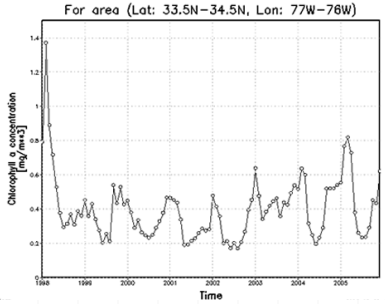
A Simple Table

Other Input Representations

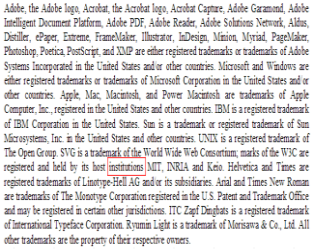
- Image, video
  - is preprocessed using Vision techniques or
  - using deep learning techniques such as deep convolutional neural networks (CNN)
- Text
  - is preprocessed using NLP techniques or
  - using deep learning techniques such as Long Short Term Memory Networks (LSTM)
- Continuous measures along time (Time series)
  - is preprocessed using Time Series analysis or
  - using deep learning techniques such as LSTMs



Image

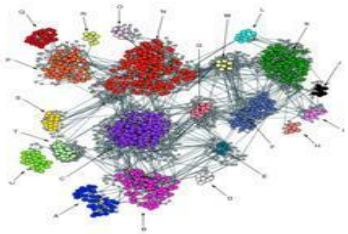


Time series



Adobe, the Adobe logo, Acrobat, the Acrobat logo, Acrobat Capture, Adobe Garamond, Adobe Intelligent Document Platform, Adobe PDF, Adobe Reader, Adobe Solutions Network, Aldus, Distiller, ePaper, Extreme, FrameMaker, Illustrator, InDesign, Minion, Myriad, PageMaker, Photoshop, Poetica, PostScript, and XMP are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States and/or other countries. Microsoft and Windows are either registered trademarks or trademarks of Microsoft Corporation in the United States and/or other countries. Apple, Mac, Macintosh, and Power Macintosh are trademarks of Apple Computer, Inc., registered in the United States and other countries. IBM is a registered trademark of IBM Corporation in the United States. Sun is a trademark or registered trademark of Sun Microsystems, Inc. in the United States and other countries. UNIX is a registered trademark of The Open Group. SVG is a trademark of the World Wide Web Consortium; marks of the W3C are registered and held by its host institutions MIT, INRIA and Keio. Helvetica and Times are registered trademarks of Linotype-Hell AG and/or its subsidiaries. Arial and Times New Roman are trademarks of The Monotype Corporation registered in the U.S. Patent and Trademark Office and may be registered in certain other jurisdictions. ITC Zapf Dingbats is a registered trademark of International Typeface Corporation. Ryumin Light is a trademark of Morisawa & Co., Ltd. All other trademarks are the property of their respective owners.

Text



Graph

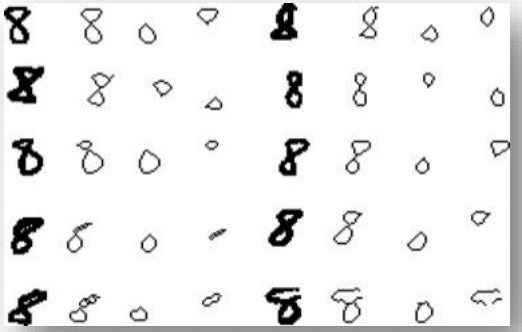
You don't Always need Machine Learning!

- Machine Learning definition (supervised):
  - The ability to learn and to improve with experience instead of using pre-determined rules.
- Consider the following two tasks:

Problem: Is **m** a prime number?

Solution: test up to $\sqrt{m}$ to see if m can be factored into two values.

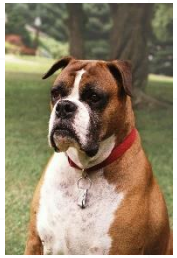
Testing for
Prime Numbers



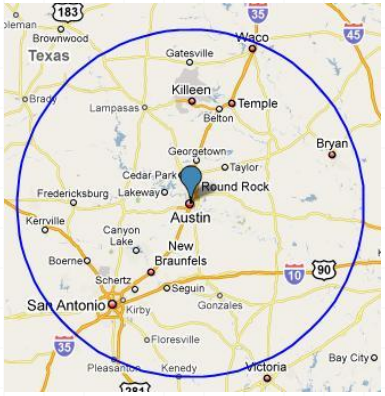
Recognizing
Handwritten Digits

Which Task Requires ML?

- Dog Recognition



- Location Proximity Detection from GPS Signal



Which Task Requires ML?

- Speech Recognition

- Detecting if a given sentence is in English or German

English: What came first: the chicken or the egg?
[wʌt keɪm fɛːrst ðə ʧɪkən ɔːr ðə ɛg]

Dutch: Wat kwam eerst: de kip of het ei?
[ʋɑt kʋam ɛːrst də kɪp ɔf hɛt ɛɪ]

German: Was kam erst: die Henne oder das Ei?
[vas kam eːɐst diː hɛnə oːdɐ das ai]

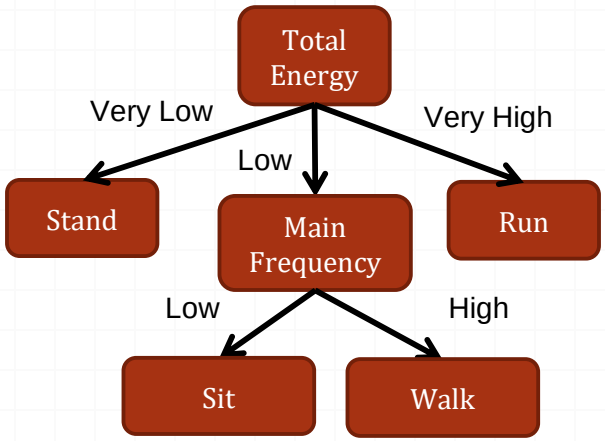
en.wikipedia.org and-bmod.com

“When” Learning is needed?

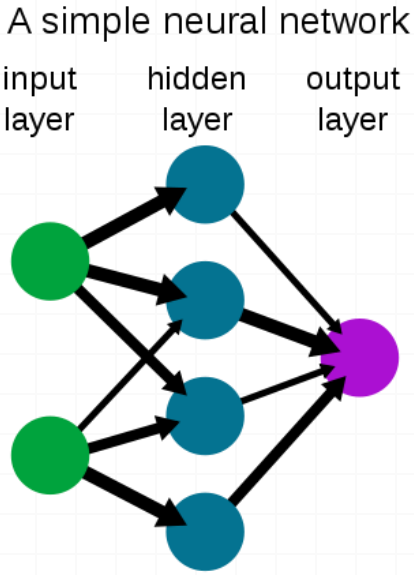
- There is no need to “learn” to calculate payroll
- Learning is used when:
  - Human expertise does not exist (navigating on Mars),
  - Humans are unable to explain their expertise (speech recognition, image analysis)
  - Solution changes in time (decision support during surgery)
  - Solution needs to be adapted to particular cases (personalized medicine)

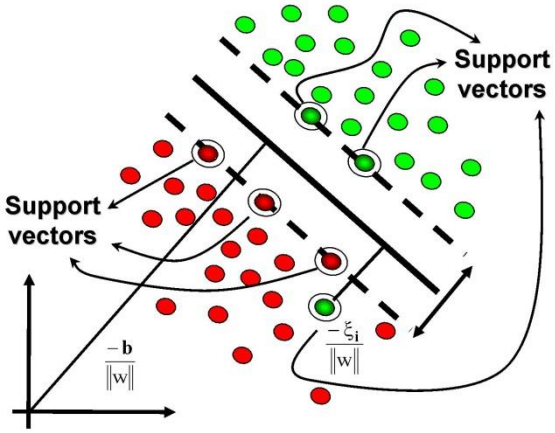
Example ML Algorithms

- Linear Regression
- Decision trees, neural network, support vector machine, …



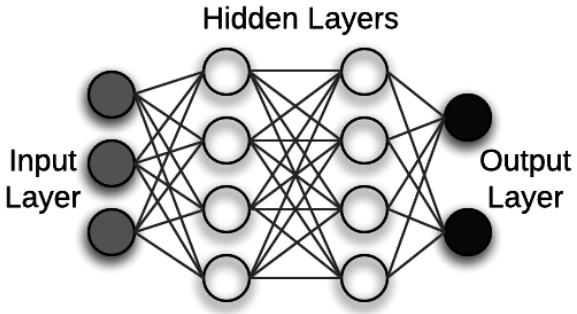
A simple decision tree





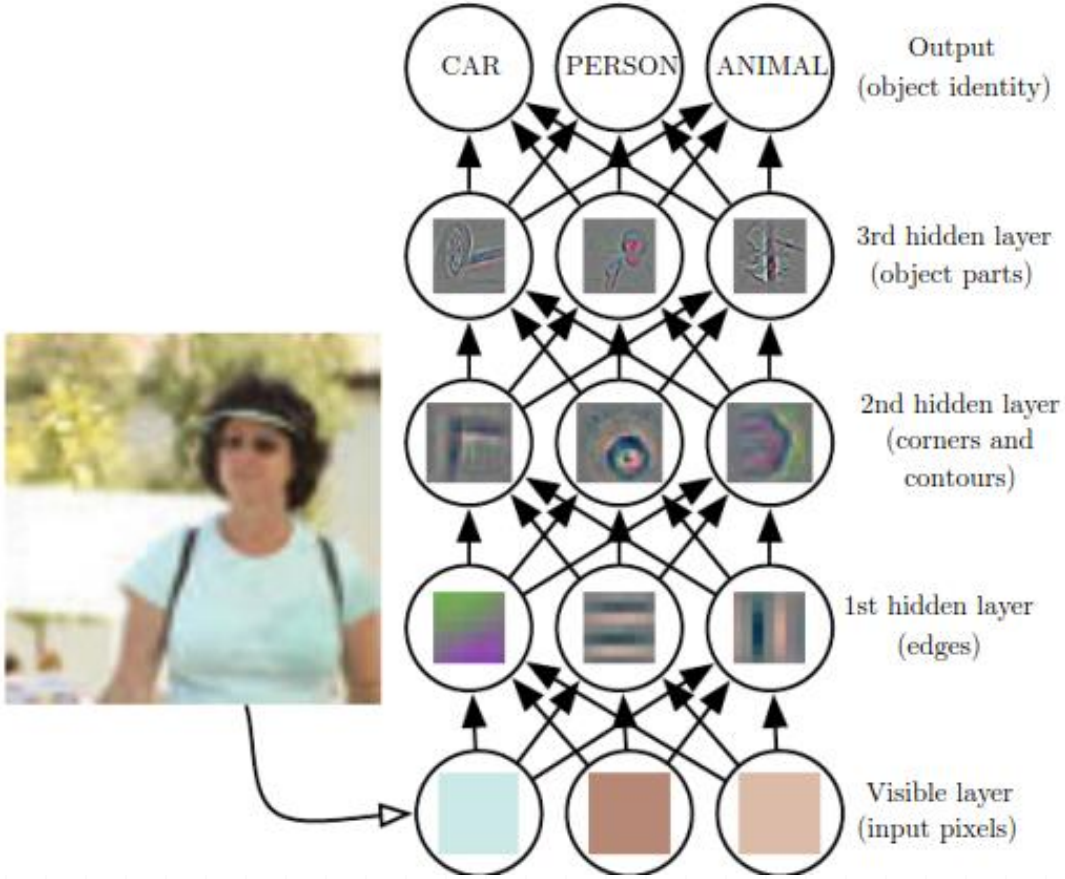
Support Vector Machines

Deep Learning



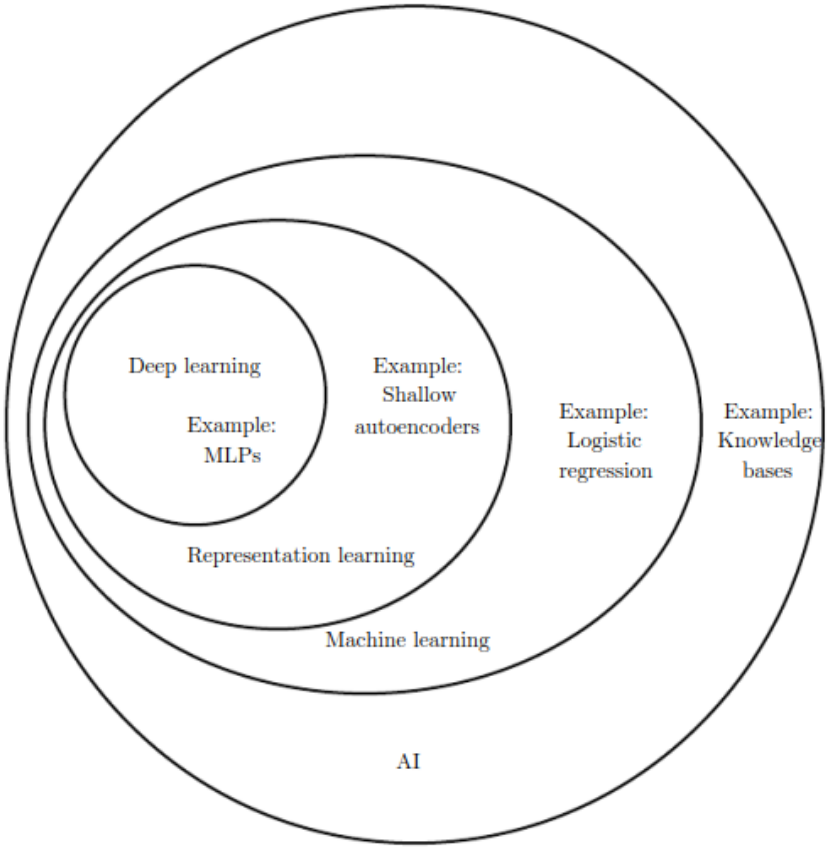
A simple Neural network

Deep Neural network



Relation with Sub-areas

- Deep learning is not equal to machine learning



ML History 1

- 1950 — Alan Turing creates the “Turing Test”
- 1952 — Arthur Samuel wrote the first computer learning program to play checkers.
- 1957 — Frank Rosenblatt designed the first neural network for computers (the perceptron)
- 1967 — The “nearest neighbor” algorithm was written

Bernard Marr, A Short History of Machine Learning -- Every Manager Should Read

ML History 2

- 1986 — back-propagation by Rumelhart
- 1992 — SVMs close to their current form introduced by Vapnik
- 1997 — LSTM introduced.
- 1997 — IBM's Deep Blue beats the world champion at chess.
- 2006 — Geoffrey Hinton coins the term "deep learning"

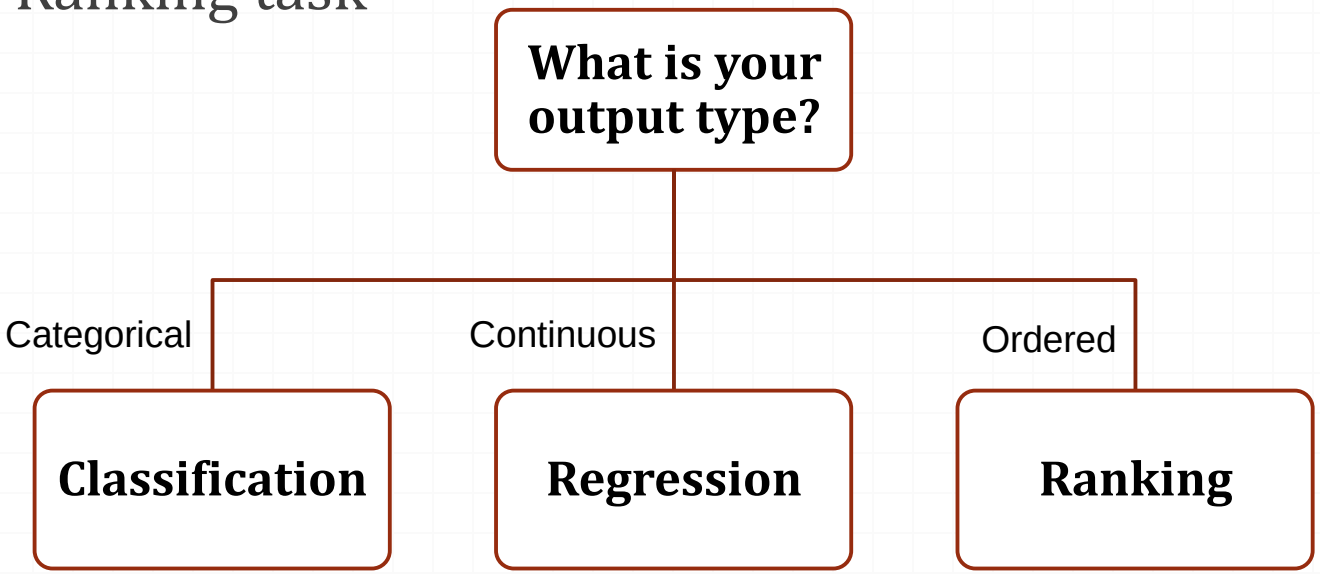
Bernard Marr, A Short History of Machine Learning -- Every Manager Should Read

ML History 3

- 2011 — IBM's Watson beats its human competitors at Jeopardy.
- 2014 – Facebook develops DeepFace
- 2016 – Google's algorithm beats a professional player at the board game Go
- 2018 - moving beyond ImageNet

Task Type

- Categorical: Classification task
  - Classifier
- Continuous: Regression task
- Ordered: Ranking task



Supervised vs. Unsupervised Learning

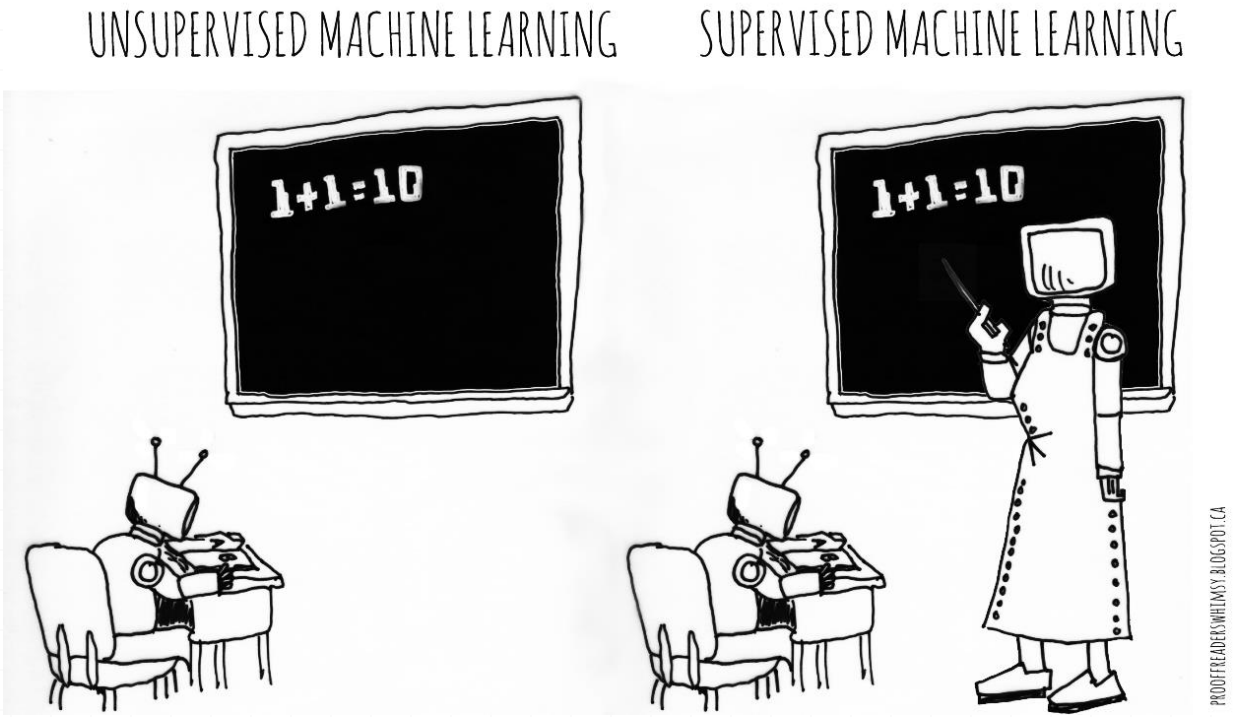
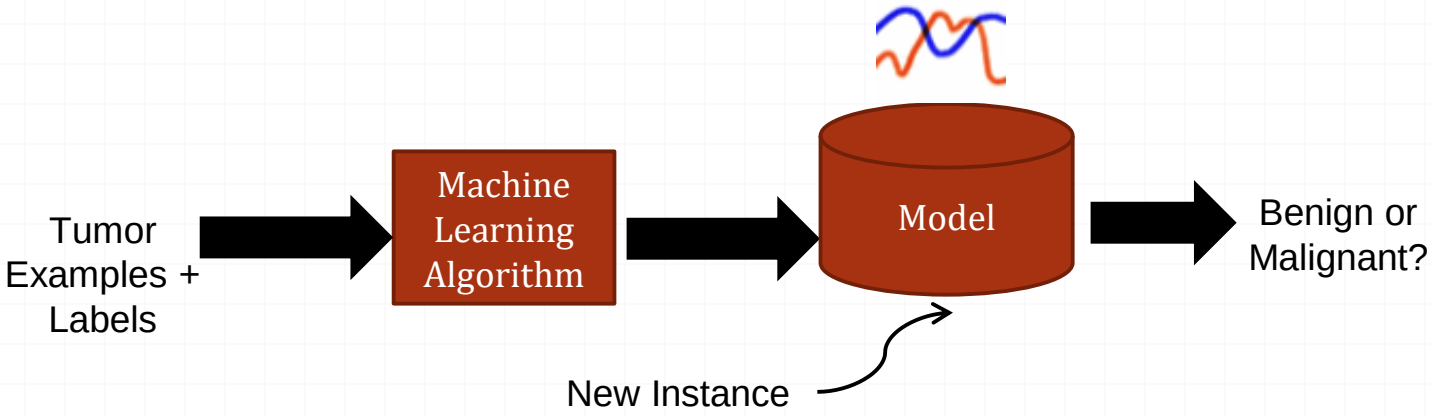


Image from *http://prooffreaderswhimsy.blogspot.com/2014/11/machine-learning.html

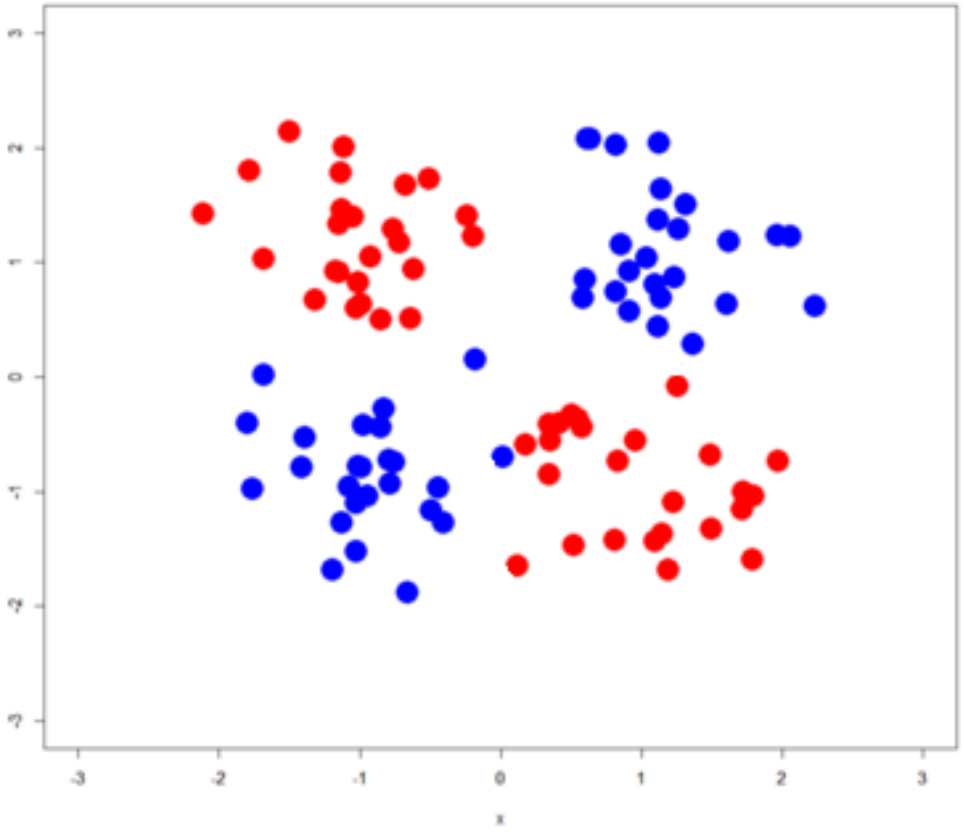
Supervised Machine Learning

- Goal is Prediction (classification or regression)
- Example:
  - **Input:** examples of benign (-) and malignant (+) tumors defined in terms of tumor shape, radius, ..
  - **Output:** predict whether a previously unseen example is benign or malignant

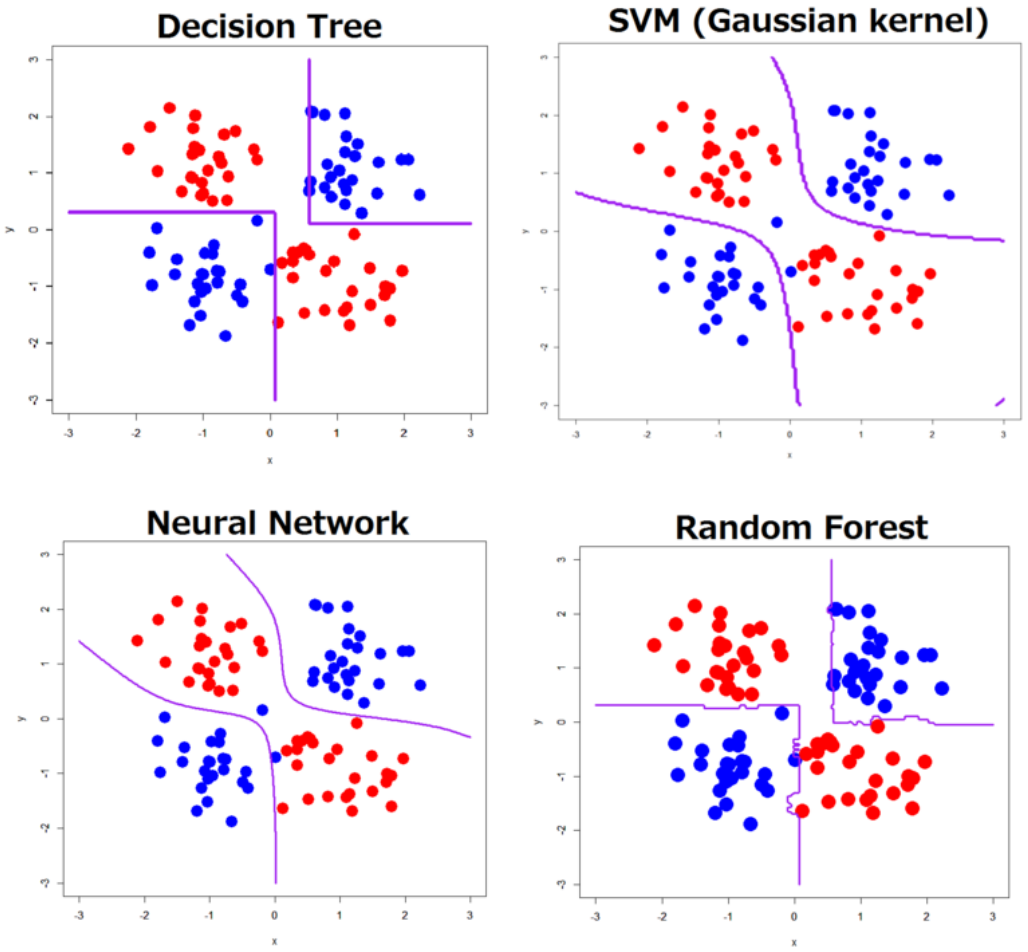


Decision Boundary

- Example dataset

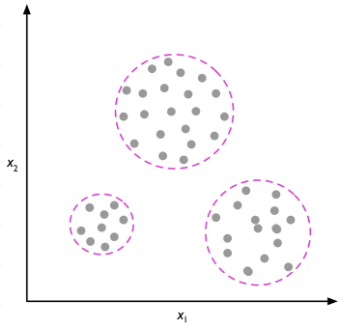


Example Decision Boundaries

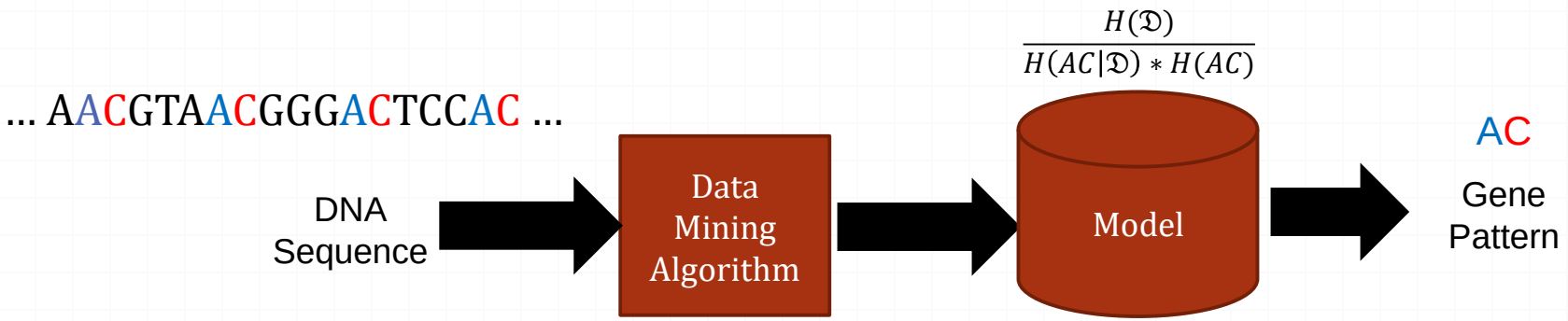


http://www.kdnuggets.com/2015/06/decision-boundaries-deep-learning-machine-learning-classifiers.html

Unsupervised Machine Learning



- Also known as data mining
- Goal is knowledge discovery
- Example:
  - **Input:** DNA Sequence as a long string of {A,C,G,T}
  - **Output:** frequent subsequences (gene patterns)



Supervised vs. Unsupervised Learning

- **Supervised Learning** ("learn from my example")
  - Goal: A program that performs a task as good as humans.
  - TASK – well defined (the target function)
  - EXPERIENCE – training data provided by a human
  - PERFORMANCE Metric – error/accuracy on the task
- **Unsupervised Learning** ("see what you can find")
  - Goal: To find some kind of structure in the data.
  - TASK – vaguely defined
  - No EXPERIENCE: no labeled data
  - No PERFORMANCE Metric (but, there are some evaluations metrics)

Beyond Supervised/Unsupervised

- Also
  - Semi-supervised learning => when a small amount of data is labeled
  - Transfer Learning => when labeled data is available in another domain