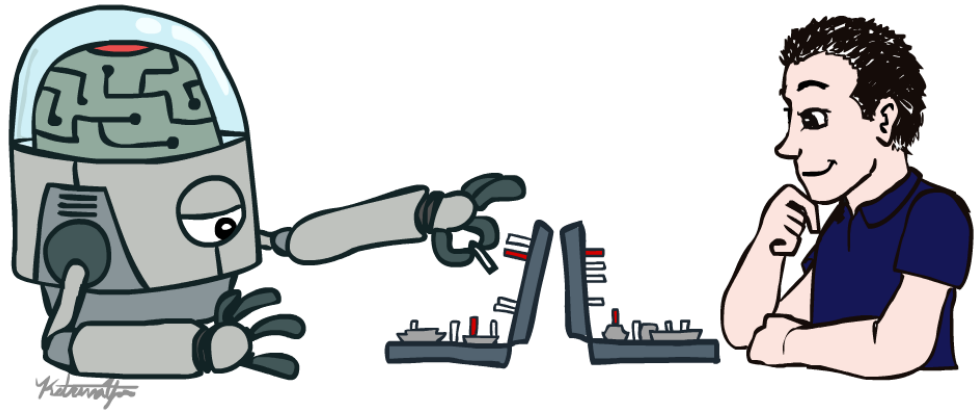
Lecture 11

Ashis Kumar Chanda
chanda@rowan.edu



RowanUniversity

Our Status in AI Course

- **Part I: Search and Planning**
- **Part II: Probabilistic Reasoning**
- **We are now on Part III: Machine Learning**
  - Classification
  - Natural language processing (NLP)
  - Computer vision
  - … lots more!

Background

- What is Machine learning (ML)?
  - The study of computer algorithms that can improve automatically through **experience** and using **data**.

- Why and when do we need to apply machine learning?
  - Ex: Automatic coffee cup Filling Machines
    - The size of coffee cup is fixed.
  - Ex: Automatic car driving system
    - Making brake system to stop car after observing 'yellow' sign.

ML Examples

Image recognition
- Label an x-ray as cancerous or not
- Assign a name to a photographed face
- Recognize handwriting

Sentiment analysis
- Understanding rating of movie review
- Identifying disaster type tweets
- Identifying patient's severity

Predictive analytics
- Weather: sunny or rainy day
- Stock market price
- Predicting whether a transaction is fraudulent or not

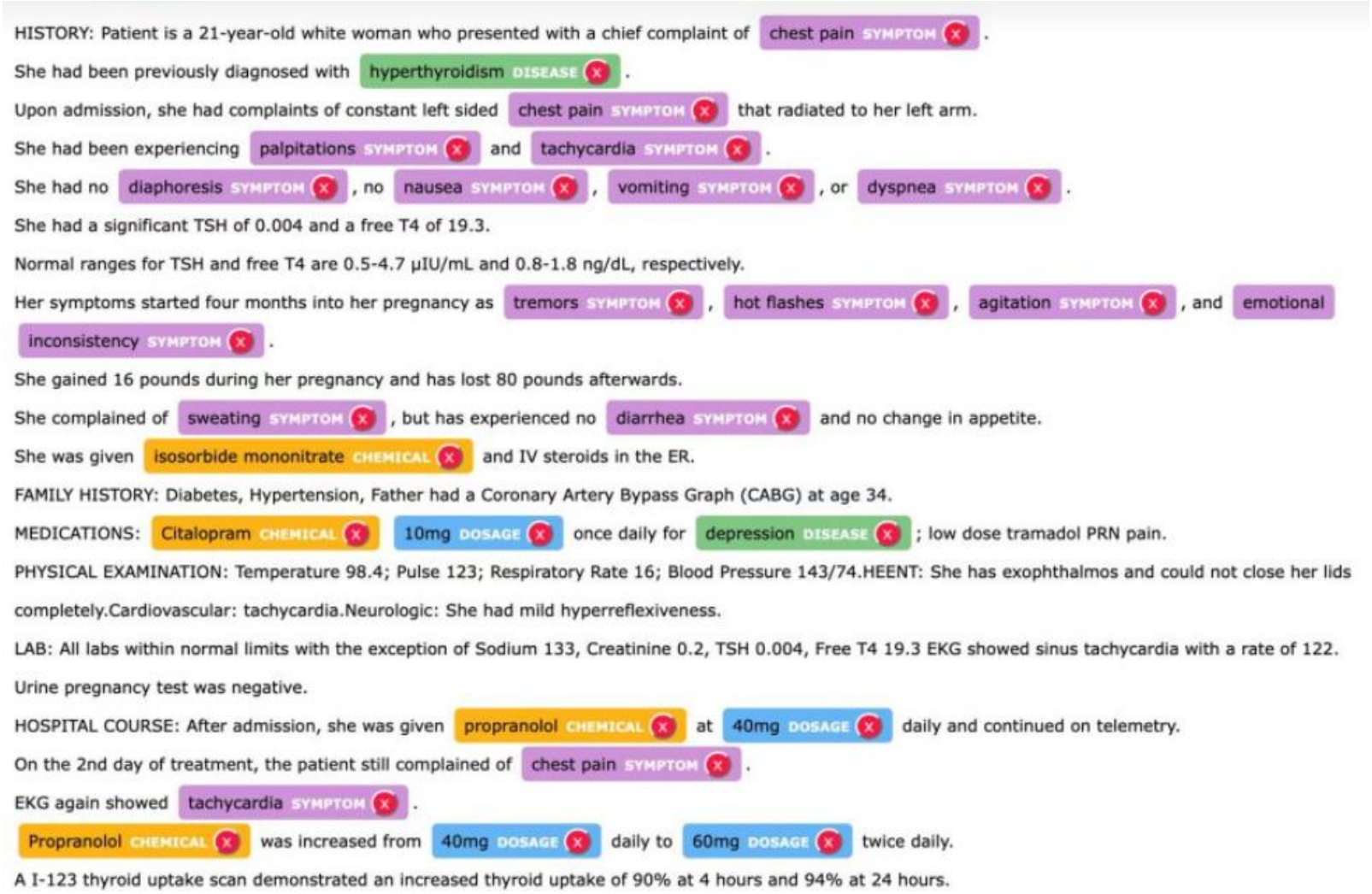
Extraction
- Parts of speech extraction from text
- Generate a model to predict vocal cord disorders
- Develop methods to prevent, diagnose, and treat the disorders

ML Examples

Let's see some real examples

https://imagerecognize.com/

ML Examples: Finding entity



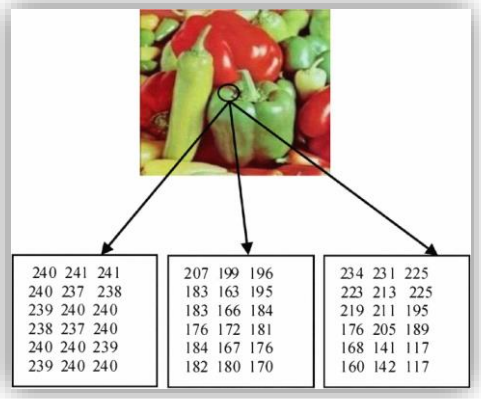
Data

- There are two types of data – 1. Structured, and 2. Un-structured data.

```
Admission Date:[**2187-12-26**] Discharge Date: [**2187-12-31**]

Date of Birth:  [**2125-11-20**]            Sex:   M

Service: CARDIOTHORACIC

Allergies:
No Known Allergies / Adverse Drug Reactions

Attending:[**First Name3 (LF) 1505**]
Chief Complaint:
Chest pain

Major Surgical or Invasive Procedure:
[**2187-12-27**] Three Vessel Coronary Artery Bypass Grafting
utilizing the left internal mammary artery to left anterior
descending, with vein grafts to the obtuse marginal and PDA.

History of Present Illness:
This is a 62 yo male with PMH signififcant for hypertension and
hypercholesterolemia. Patient admits to experiencing chest
tightness with left hand numbness and diaphoresis for the first
time 4 days prior to admission while carrying a load up a flight
of stairs. The chest pain was relieved with ASA after 20
```

Unstructured data

Income	Age	Student	Loan
Low	Young	No	N
Low	Young	Yes	N
High	Senior	No	Y
High	Senior	Yes	Y
High	Young	Yes	Y
Low	Young	No	Y
Low	Senior	No	N



Picture data is stored in RGB format

Structured data

Data Examples

Let's see some real life datasets

https://archive.ics.uci.edu/ml/datasets.php
https://www.kaggle.com/datasets?fileType=csv

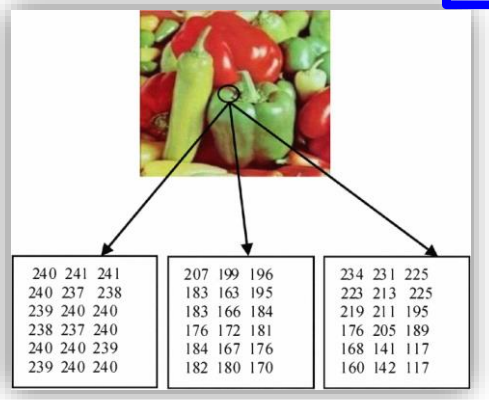
Learning methods

- There are two types of learning methods – 1. Supervised, and 2. Unsupervised.

```
Admission Date:[**2187-12-26**] Discharge Date: [**2187-12-31**]

Date of Birth:  [**2125-11-20**]            Sex:   M

Service: CARDIOTHORACIC

Allergies:
No Known Allergies / Adverse Drug Reactions

Attending:[**First Name3 (LF) 1505**]
Chief Complaint:
Chest pain

Major Surgical or Invasive Procedure:
[**2187-12-27**] Three Vessel Coronary Artery Bypass Grafting
utilizing the left internal mammary artery to left anterior
descending, with vein grafts to the obtuse marginal and PDA.

History of Present Illness:
This is a 62 yo male with PMH signififcant for hypertension and
hypercholesterolemia. Patient admits to experiencing chest
tightness with left hand numbness and diaphoresis for the first
time 4 days prior to admission while carrying a load up a flight
of stairs. The chest pain was relieved with ASA after 20
```

No label for patient history data

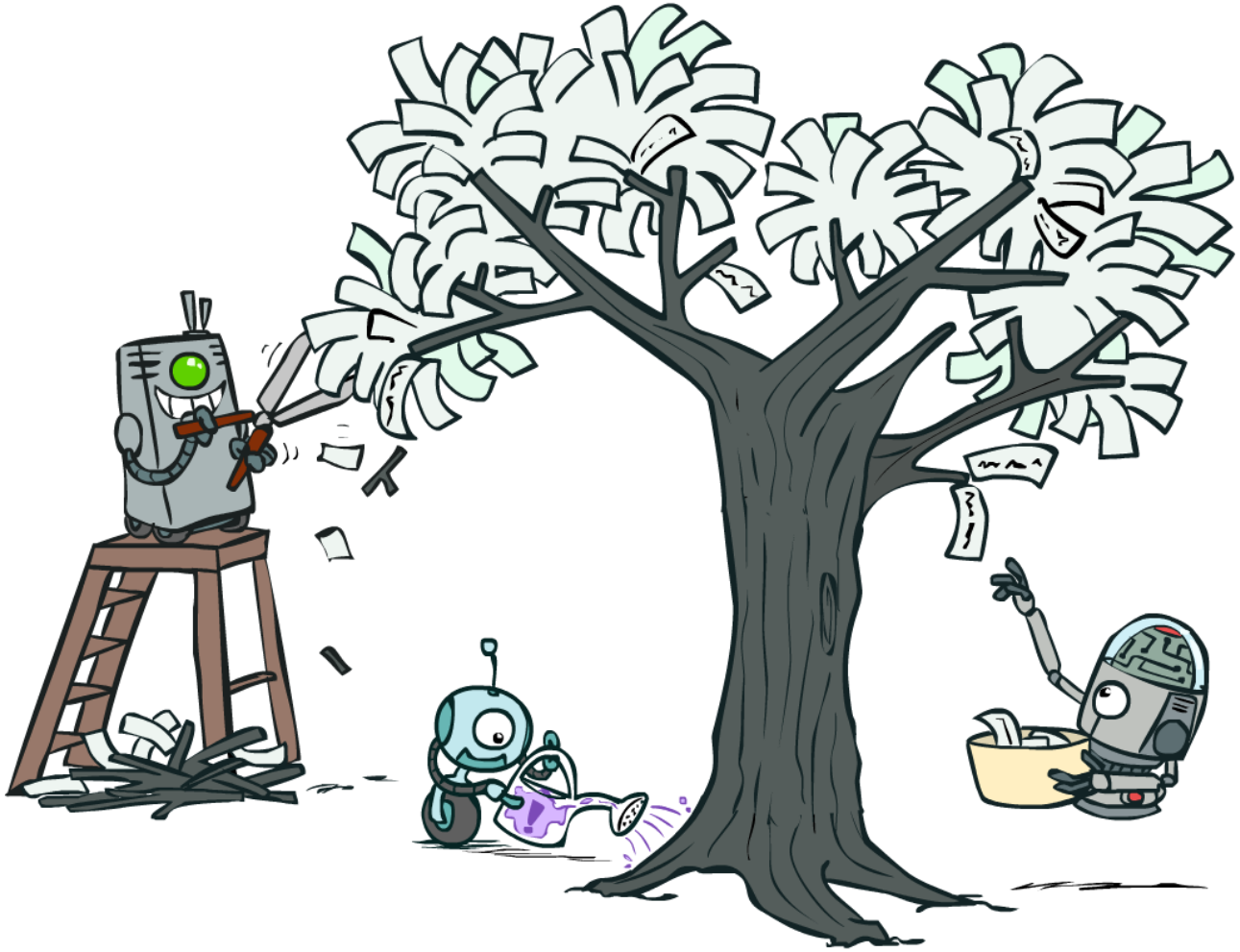
Income	Age	Student	Loan
Low	Young	No	N
Low	Young	Yes	N
High	Senior	No	Y
High	Senior	Yes	Y
High	Young	Yes	Y
Low	Young	No	Y
Low	Senior	No	N



Fruit

Let’s learn some ML methods

Decision Trees



Choosing an Attribute

- Idea: a good attribute splits the examples into subsets that are (ideally) "all positive" or "all negative"

Income	Age	Student	Loan
Low	Young	No	N
Low	Young	Yes	N
High	Senior	No	Y
High	Senior	Yes	Y
High	Young	Yes	Y
Low	Young	No	Y
Low	Senior	No	N

- So: we need a measure of how "good" a split is, even if the results aren't perfectly separated out

Sample Data

- Try to predict loan for the following dataset.
- It looks like an “OR” relationship.

High income	Senior age	Loan
No	No	No
Yes	No	Yes
No	Yes	Yes
Yes	Yes	Yes

Sample Data

- Design a decision tree for the following data set to predict loan.
- Note: Marital status has three types of values.

young age	Marital status	Loan
No	Single	No
No	Single	No
No	Married	No
No	Married	No
No	Single	Yes
No	Single	Yes
No	Divorce	Yes
Yes	Divorce	No
Yes	Married	No
Yes	Married	No

Quiz

- Design a decision tree for the following data set to predict loan.

Income	Age	Student	Loan
Low	Young	No	N
Low	Young	Yes	N
High	Senior	No	Y
High	Senior	Yes	Y
High	Young	Yes	Y
Low	Young	No	Y
Low	Senior	No	N

Entropy

- ID3 (Iterative Dichotomiser 3) is an algorithm invented by Ross Quinlan used to generate a decision tree from a dataset.

- Picking an attribute as a root that has high information gain.
  - Information is expected:

$$H(\langle p_1, \ldots, p_n \rangle) = \sum_{i=1}^{n} -p_i \log_2 p_i$$

- Also called the entropy of the distribution.

Note on Decision Tree

- Tree depth is limited by the number of attributes.
- DTree may result local minimum.
- DTree is an **offline** method.
  - What is offline and online method?
  - Offline method: If a trained model starts to train again for a new data.
  - Online method: When a trained model can decide the label of a new data without retraining the model.
- DTree is also known as **inductive learning** method, because it checks each attribute one by one to make decision.

Random forest

- The random forest is a classification algorithm consisting of many trained tree to avoid overfitting.
- It randomly takes sqrt(A) number of attributes from **A** attributes to build a decision treen.
- Continues the process **k** times to build **k number of tree**.
- Finally, it takes **vote** among k trained tree to select the maximum voted class as a label for an example or new data.

Splitting dataset

- **Randomly** split your dataset into three separate files,
  - 1) train, 2) validation, 3) test dataset.
- The splitting ratio:
  - 80%, 10%, and 10% for train, validation, and test dataset, respectively.
    - When you have a good number of data
  - 98%, 1%, and 1% for train, validation, and test dataset, respectively.
    - When you have very large data
  - Apply **k-fold cross validation** for small data

Model accuracy

- We need a metric to say how good is our trained model.

$$Accuracy\ (Acc) = \frac{TP + TN}{(TP + TN + FP + FN)}$$

$$Recall\ (True\ positive\ rate) = \frac{TP}{TP + FN}$$

$$Precision\ (positive\ predictive\ value) = \frac{TP}{TP + FP}$$

$$F1\ score = \frac{2\ *Recall\ *\ Precision}{(Recall + Precision)}$$

Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Model accuracy

- We need a metric to say how good is our trained model.
- We also need to figure out that the trained model is useful or not.
  - Does it work better than any other old or simple methods?
- **Baseline method**:
  - Any simple or logical method to determine the class accuracy.
  - For real research, usually use **previous work** as a (strong) baseline.

Baselines

- First step: get a baseline
  - Baselines are very simple "straw man" procedures
  - Help determine how hard the task is
  - Help know what a "good" accuracy is

- Weak baseline: most frequent label classifier
  - Gives all test instances whatever label was most common in the training set
  - E.g., for spam filtering, we have only 25% spam email.
  - So, if we label everything as ham, we will have 75% accuracy.
  - Now, a trained classifier that gets 70% isn't very good…

- For real research, usually use previous work as a (strong) baseline.

Overfitting

- **Overfitting**: Good performance on the training data, poor generalization to other data.

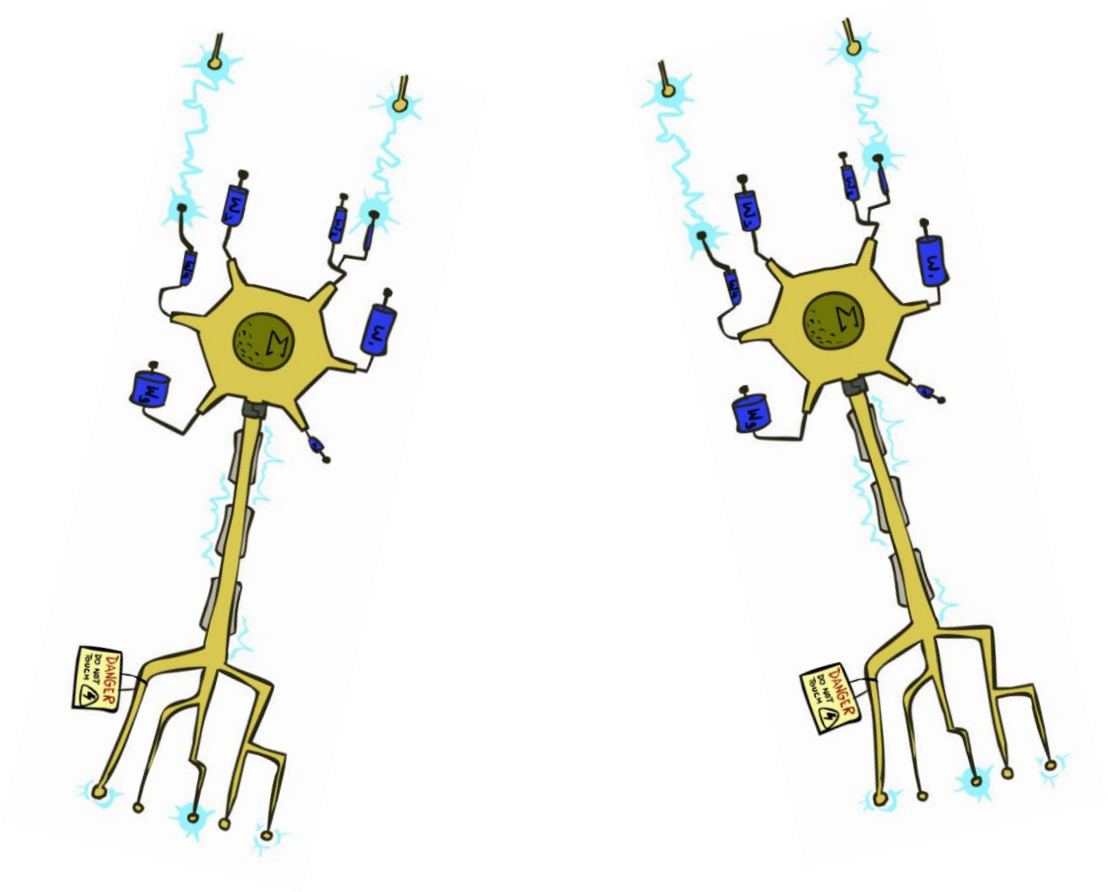
- **Techniques to reduce overfitting:**
  - Increase training data.
  - Reduce model complexity.

- **Underfitting** is the case where the model has " not learned enough" from the training data, resulting in low generalization and unreliable predictions.

- **Techniques to reduce underfitting:**
  - Increase model complexity
  - Increase the number of features
  - Remove noise from the data.
  - Increase the number of epochs

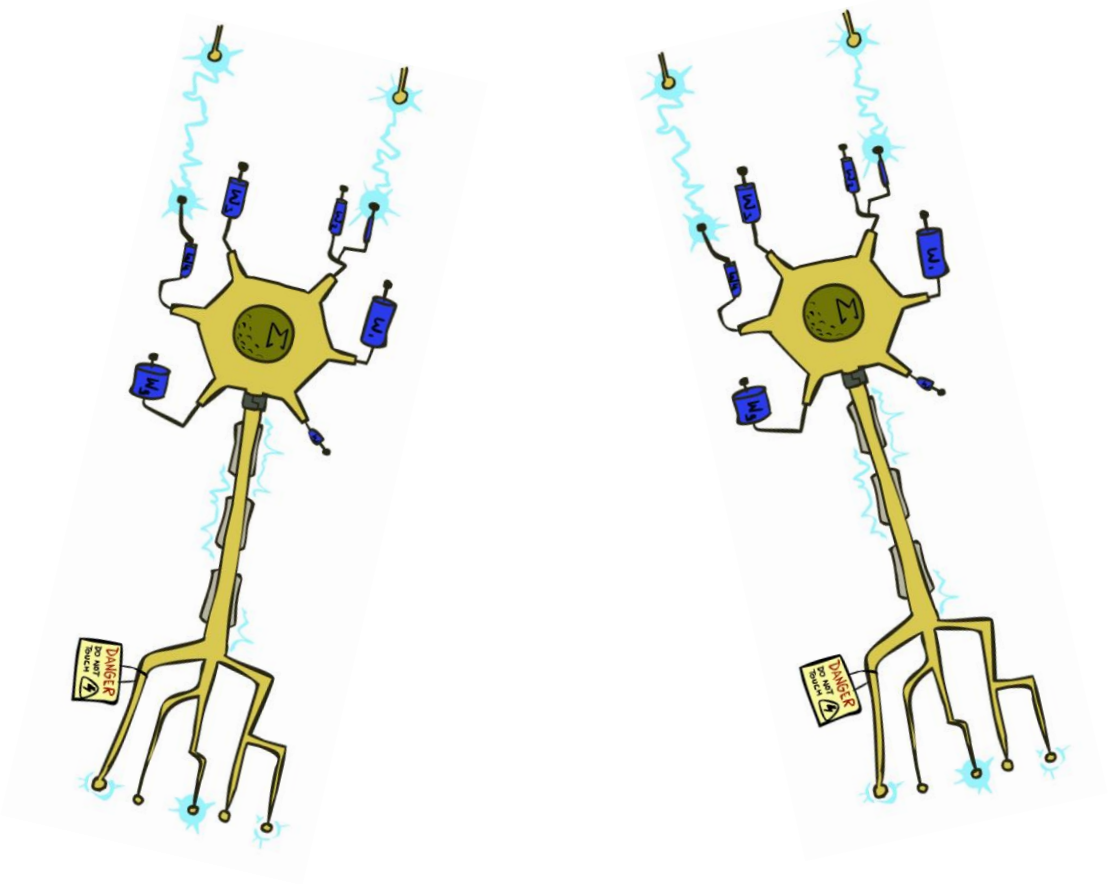
Today

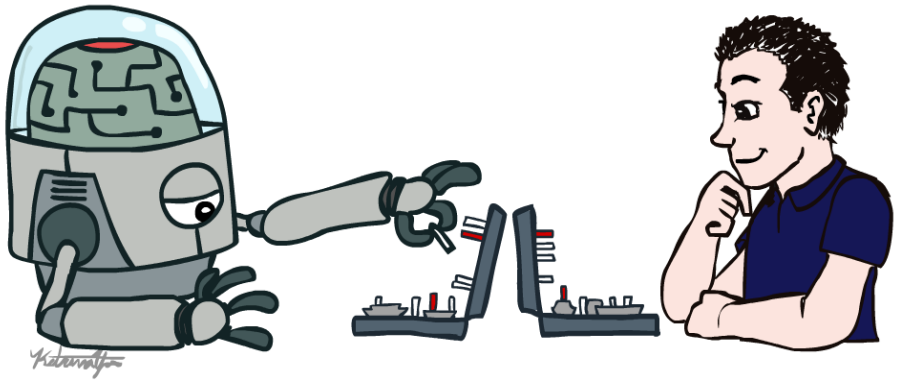
- Machine learning
- Decision tree
- Random forest
- Train-test dataset
- K-fold cross validation
- Confusion matrix
- Baselines
- Overfitting



Next class

- Linear regression
- Neural network
- Backpropagation





Thanks!