Machine Learning in Healthcare

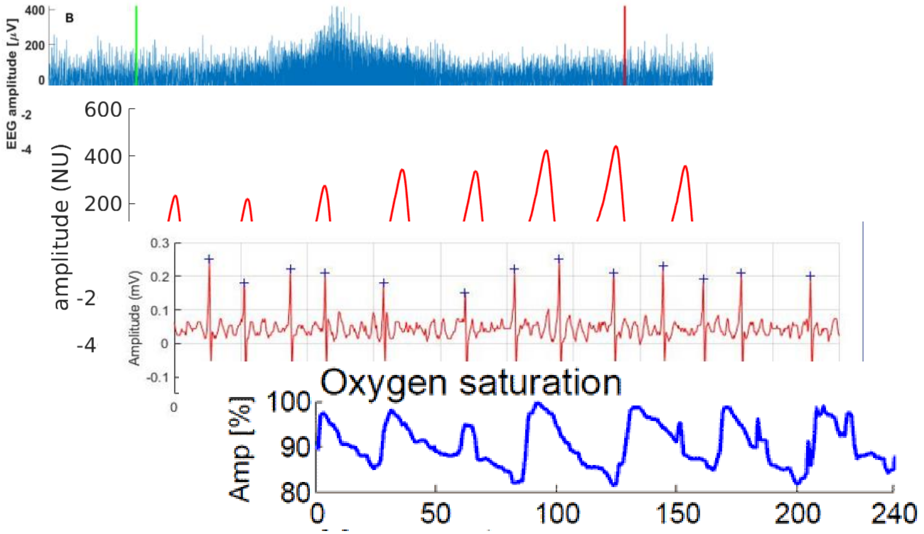
#L01-Introduction to machine learning in healthcare

Technion-IIT, Haifa, Israel

Asst. Prof. Joachim Behar
Biomedical Engineering Faculty, Technion-IIT
Artificial intelligence in medicine laboratory (AIMLab.)
https://aim-lab.github.io/
Twitter: @lab_aim

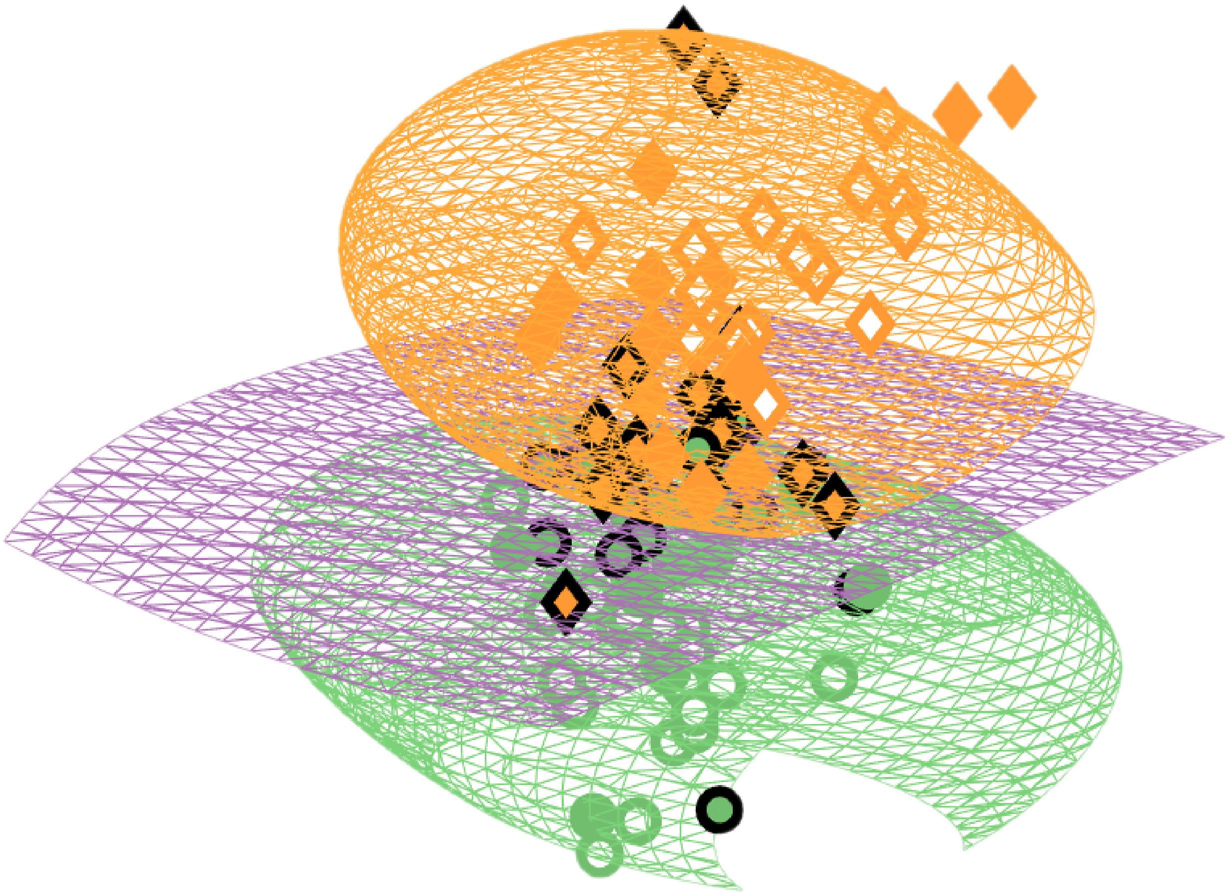
AIMLab. research

- Researches innovative **pattern recognition algorithms** to exploit the **information encrypted** within large datasets of **physiological time series**.
  - Engineer novel intelligent remote patient monitoring systems.
  - Fundamental medical research.



- **Running projects:**
  - Sleep medicine: from sleep medicine to medicine during sleep.
  - Cardiology: atrial fibrillation diagnosis and phenotyping.
  - Coronaviruses: COVID-19.
  - OBGY: fetal electrocardiography.

Machine learning



Course structure

Course Structure - Overview

AIMLab.

- **Lecturer** – Asst. Prof. Joachim A. Behar (jbehar@technion.ac.il)
  - Mathematical background and intuition.
  - Lectures: 10.30 am - 12.30 pm on Sunday.
- Laboratory sessions and assignments:
  - **Head TA:** Moran Davoodi (smorandv@campus.technion.ac.il)
  - **TA + Assignments:** Yuval Ben Sason (yuval.b@campus.technion.ac.il)
  - **Assignments:** Kevin Kotzen (kkotzen@campus.technion.ac.il)
  - Computer labs: 4.30 - 7.30 pm on Sunday or 5.30 - 8.30 pm on Tuesday.
- Use the Forum on Moodle for questions.

Course Aims

- You will learn:
  - Python for biomedical data science.
  - Main classifiers, intuition and mathematical background.
  - Neural networks and deep learning.
  - Performance statistics in healthcare.
  - ML for diagnosis, prognosis and treatment.
  - Ground truth in medical data science.

Course prerequisites

- Prerequisites:
    - Algebra.
    - Introduction to probability.
    - Coding skills.
    - Signals and Systems.
- You should also come to this course prepared following the instructions your received in setting your environments.

Course Structure

- Three modules, 4-weeks each.

ML Foundation
Weeks 1-4

Popular classifiers
Weeks 5-8

Introduction to deep learning
Weeks 9-13

- One computational assignment for each part (3 in totals).
- One theoretical assignment.

Course Evaluation

- Assignments, each is 25% of the mark:
  - 03/11/20 - 08/12/20: First assignment released and deadline.
  - 08/12/20 - 05/01/21: Second assignment release and deadline.
  - 13 - 14/01/21: Time limited theoretical and computing assignment - <u>Individual</u>.
  - 17/01/21 - 28/02/21: Fourth assignment release and deadline.
- Other notes on dates:
  - 06/12/21 (group A) and 08/12/21 (group B): First theoretical tutorial on supervised learning.
  - 10/01/21 (group A) and 12/01/21 (group B): Second theoretical tutorial.

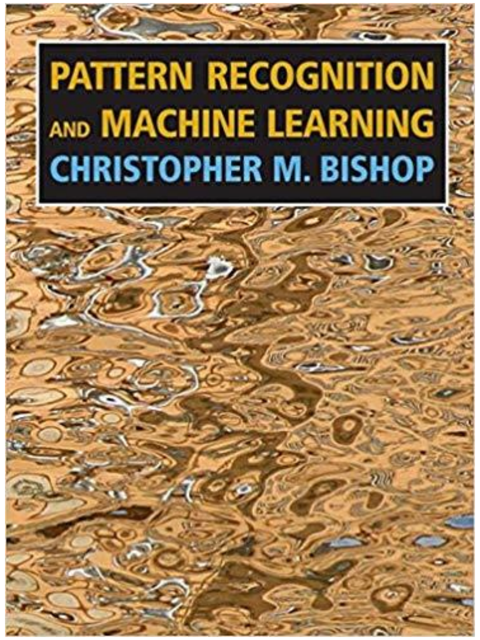
Course Evaluation

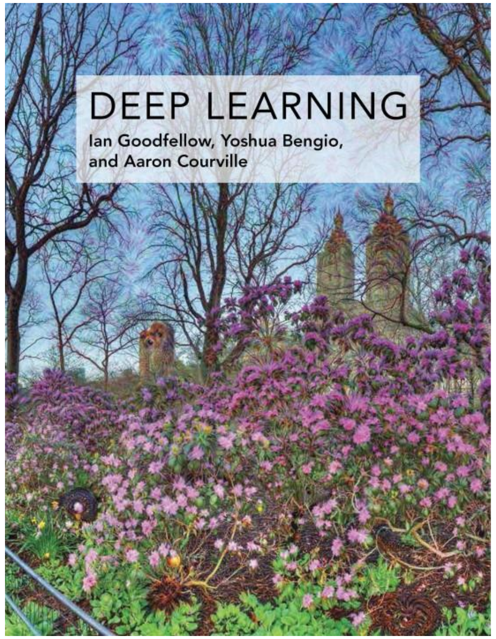
AIMLab.

- Penalties
  - Plagiarism ➔ 0%.
  - Late submission ➔ -25% per week passed the deadline (cumulative every week.)
- Important: assignments source codes are submitted through GitHub Classroom. Reports are submitted in **PDF** documents to **Yuval**.

Textbooks

- Recommended but not required for completing the course:
  - Bishop, Christopher M. Pattern recognition and machine learning. springer, 2006.
  - Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. Deep learning. MIT press, 2016.





Other resources

- List of other recommended resources:
  - Machine Learning Mastery: https://machinelearningmastery.com/
  - Toward Data Science: https://towardsdatascience.com/
  - Kaggle: https://www.kaggle.com/
  - Google: https://developers.google.com/machine-learning/crash-course/ml-intro
  - Coursera: all courses by deeplearning.ai

Introduction to ML

“The fourth industrial revolution”, why?

$$\Psi(1s) = 2a_0^{-1.5} e^{\frac{-r}{a_0}}$$

$$\Psi(2s) = \frac{1}{\sqrt{8}} a_0^{-1.5} (2 - \frac{r}{a_0}) e^{\frac{-r}{2a_0}}$$

$$\Psi(2p) = \frac{1}{\sqrt{24}} a_0^{-1.5} (\frac{r}{a_0}) e^{\frac{-r}{2a_0}}$$

Computer → “easy one!”

Human → “Give me a couple of hours…”

Computer → “not sure…”

Human → “It’s obviously a cat”

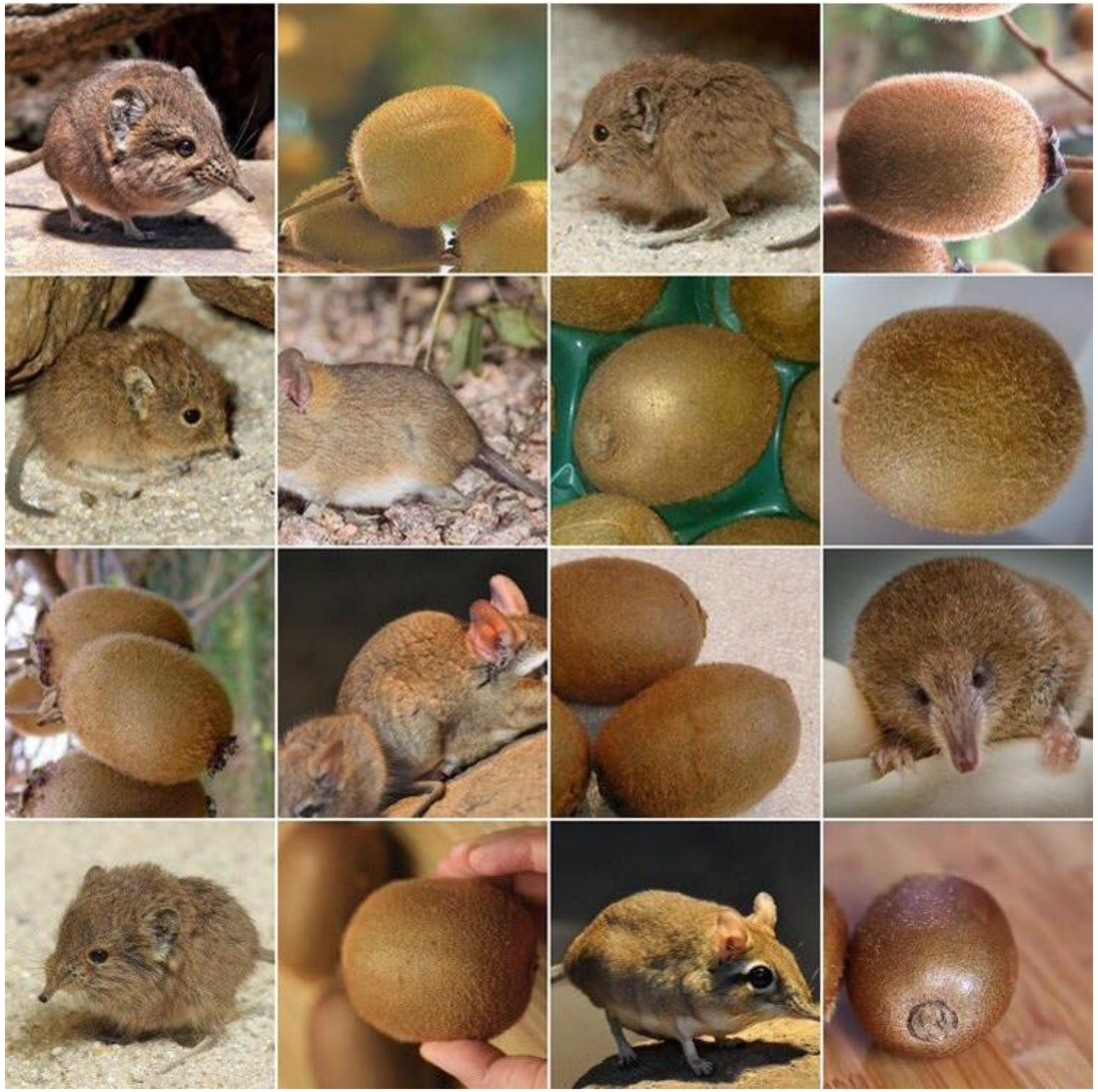
Puppy or bagel?

Credits Karen Zack

Sheepdog or mop?

Credits Karen Zack

Shrew or kiwi?



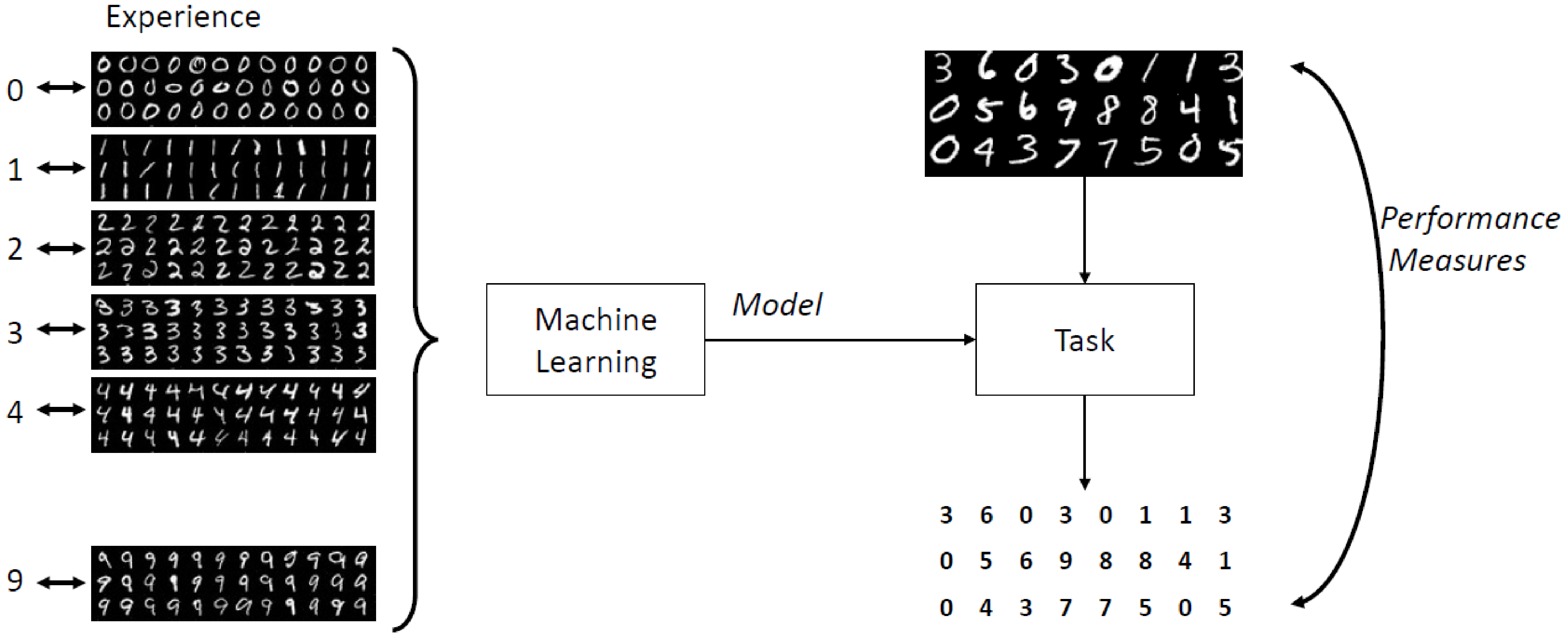
Credits Karen Zack

What is machine “learning”?

AIMLab.

- Machines have been good (better than humans!) at doing complex tasks for a long time.
- The challenge is in extracting **semantic information** from the signal.
  - Tasks that are easy for people to perform but hard to describe formally.
  - The “meaning” of an image for example, recognizing spoken words.
- How can we program a computer to extract this information?
  - This is what we will cover in this course!
  - With a particular emphasis on dealing with medical data.
- A definition of ML: “Field of study that gives computers the ability to learn without being explicitly programmed” **Arthur Samuel** (1959)

Supervised learning on the MNIST dataset



Objects Recognition: Supervised Learning

Redmon, Joseph, and Ali Farhadi. "Yolov3: An incremental improvement." arXiv preprint arXiv:1804.02767 (2018).

Playing Atari: Reinforcement Learning

Mnih, Volodymyr, et al. "Playing atari with deep reinforcement learning." arXiv preprint arXiv:1312.5602 (2013).

AlphaGo

AIMLab.

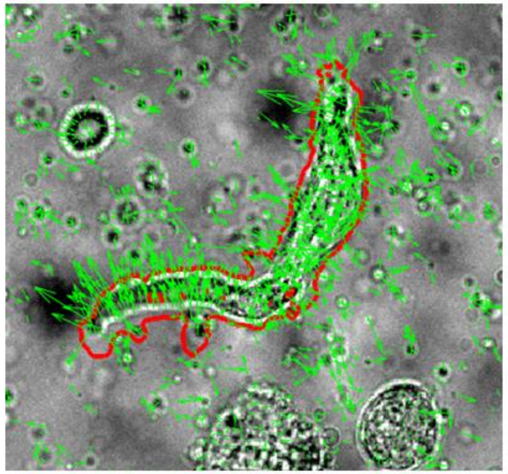
https://www.youtube.com/watch?v=WXuK6gekU1Y&ab_channel=DeepMind

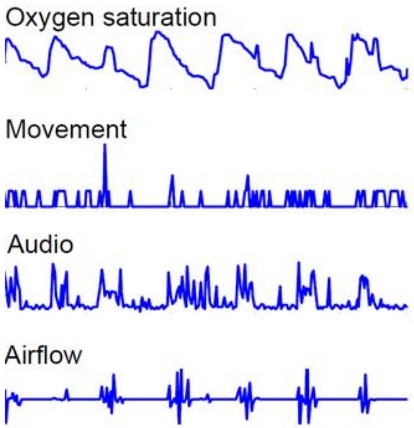
Latest news, AlphaGo Zero: https://deepmind.com/blog/alphago-zero-learning-scratch/

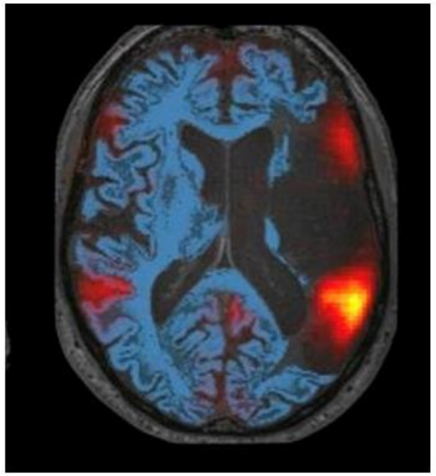
ML in Healthcare

ML in Biomedical Engineering

- Complex physiological time series analysis (e.g. ECG, EEG).
- Medical imaging (e.g. CT, MRI, ULS).
- Genetics and genomics research.
- Drug discovery.

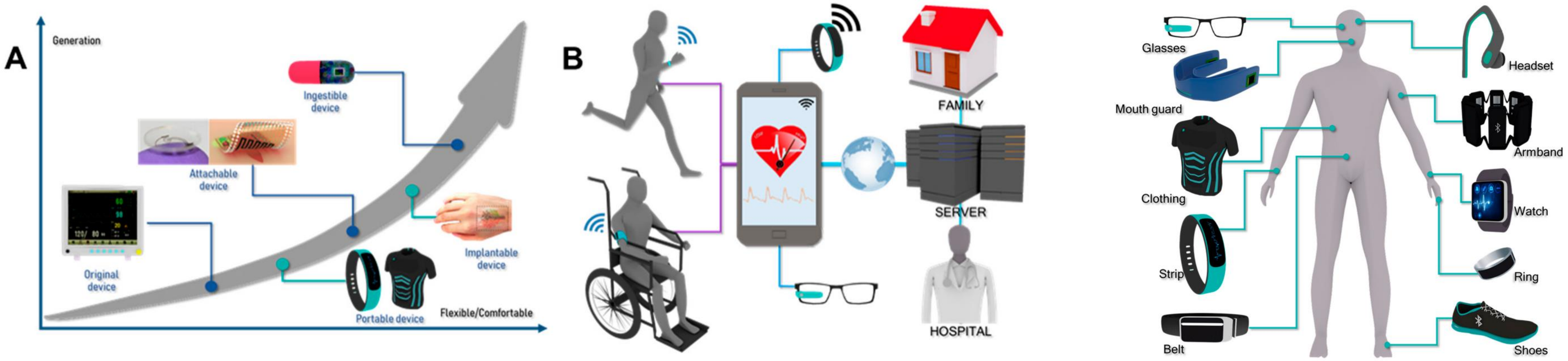






21th Century: big data in healthcare

- Easier to collect medical data in general.
- New portable sensors: open to long term monitoring and phenotyping of patients.
  - Many are low cost and widely available.



Guk, Kyeonghye, et al. "Evolution of Wearable Devices with Real-Time Disease Monitoring for Personalized Healthcare." Nanomaterials 9.6 (2019): 813.

21th Century: electronic medical record (EMR)

AIMLab.

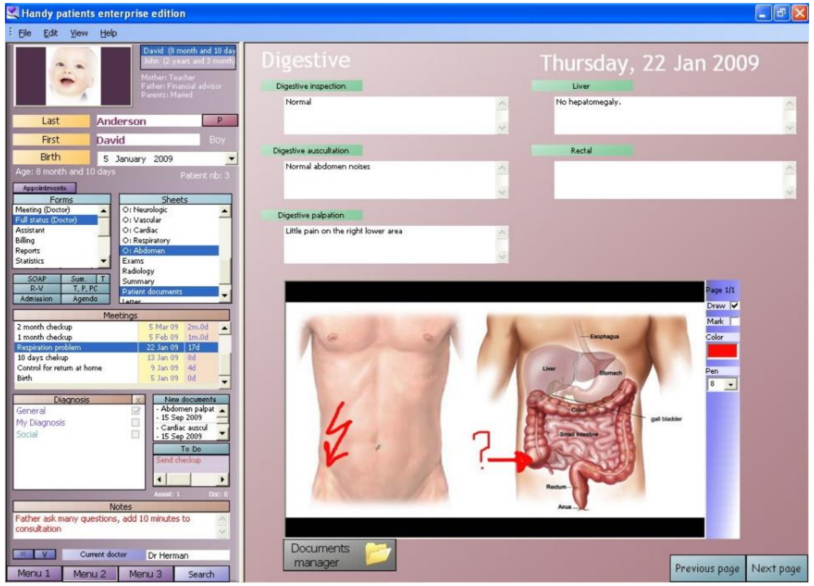
- Systematized collection of patient and population electronically-stored health information in a digital format (Gunter et al. 2005).
- Remove the need to document history in paper and is more robust, complete, standardized and accessible. SQL
- Include different types of data: demographics, medication and allergies, immunization status, laboratory test results, radiology images, vital signs, etc.

21th Century: electronic medical record (EMR)

AIMLab.

· HIPAA

- Healthcare providers use them to improve their care management.
- In term of research, they significantly ease population based study by enable to look at long term trends and changes in patients.



https://cov19-resource.om/

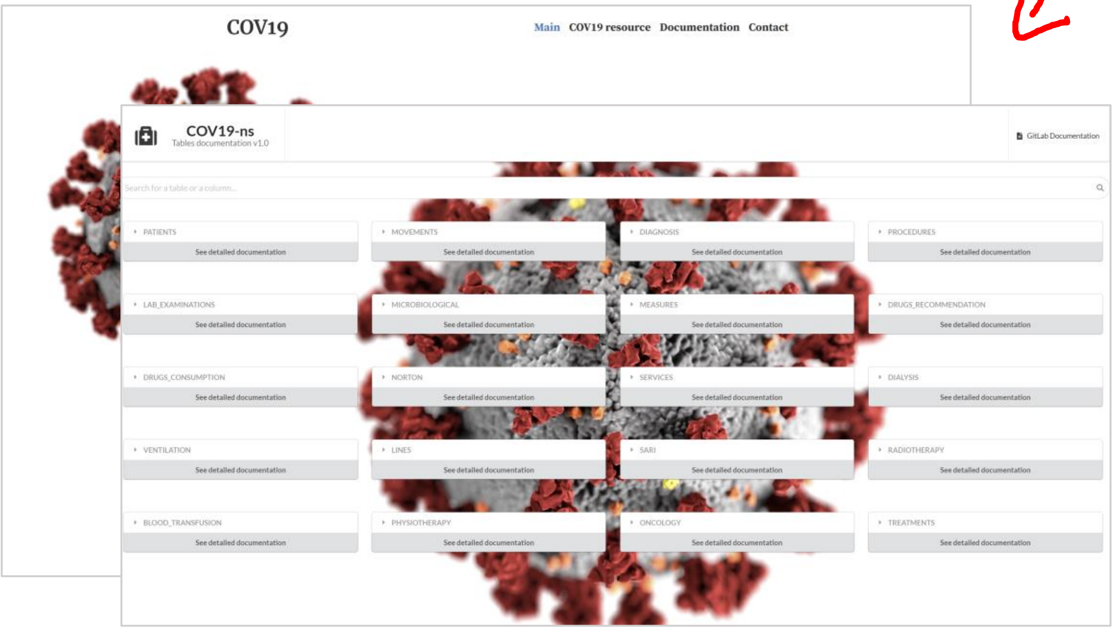
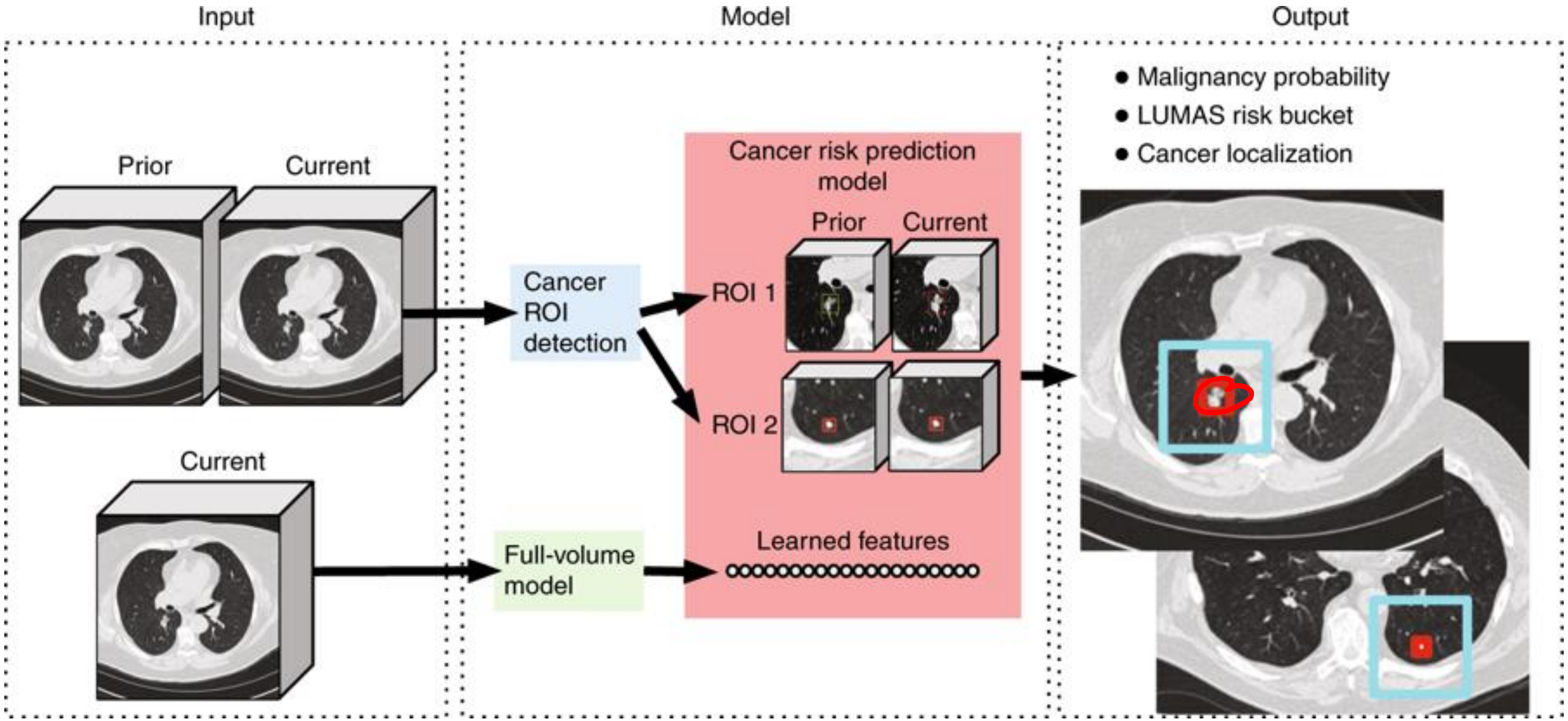


Image: By Oguntoye patients electronic medical record (free open source version), GPL, https://commons.wikimedia.org/w/index.php?curid=8894074

Examples: medical images

- Medical imaging:



Ardila, Diego, et al. "End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography." Nature medicine 25.6 (2019): 954.

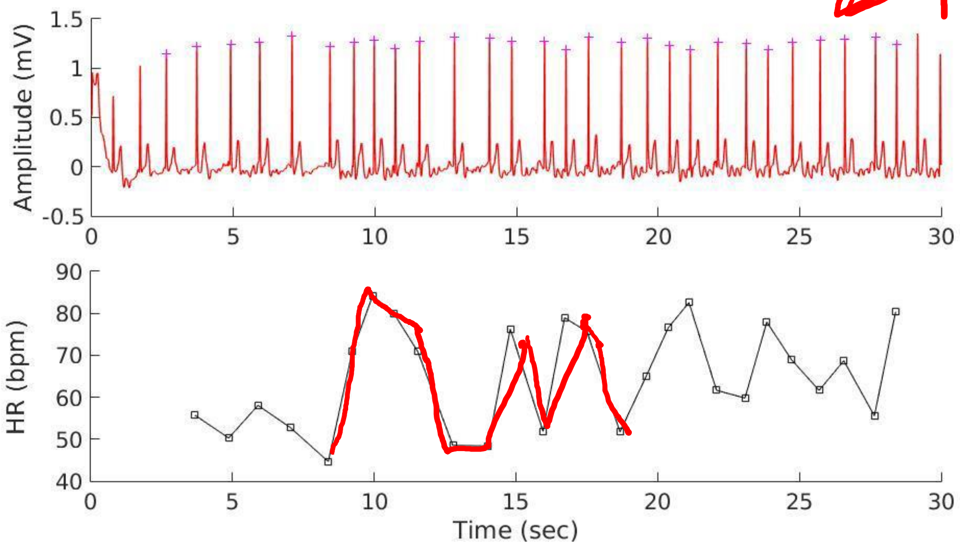
Examples: atrial fibrillation diagnosis

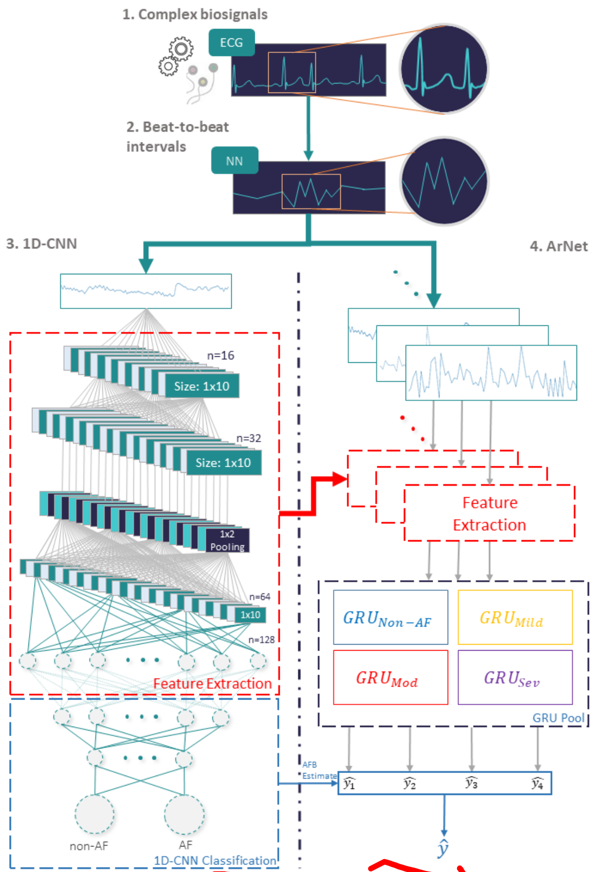
AIMLab.

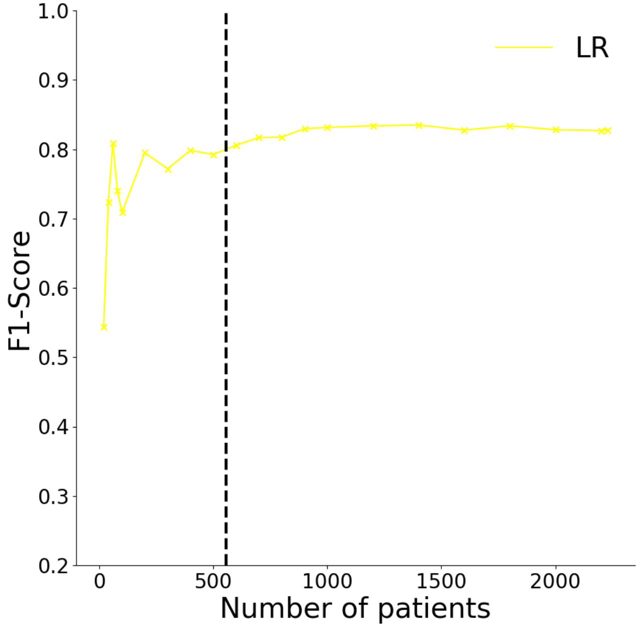
- Recognition of cardiac arrhythmias.

ECG

AF







AF

non-AF

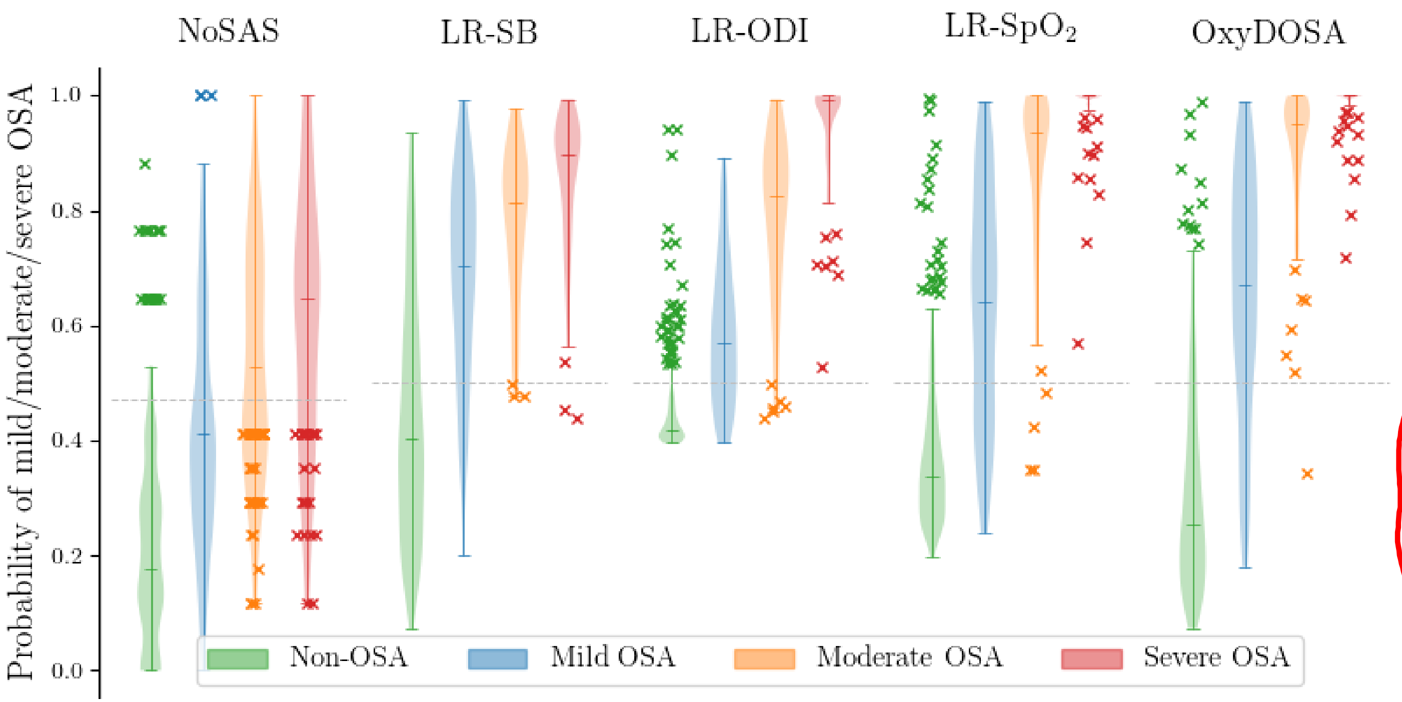
Chocron A. et al. under review in IEEE TBME

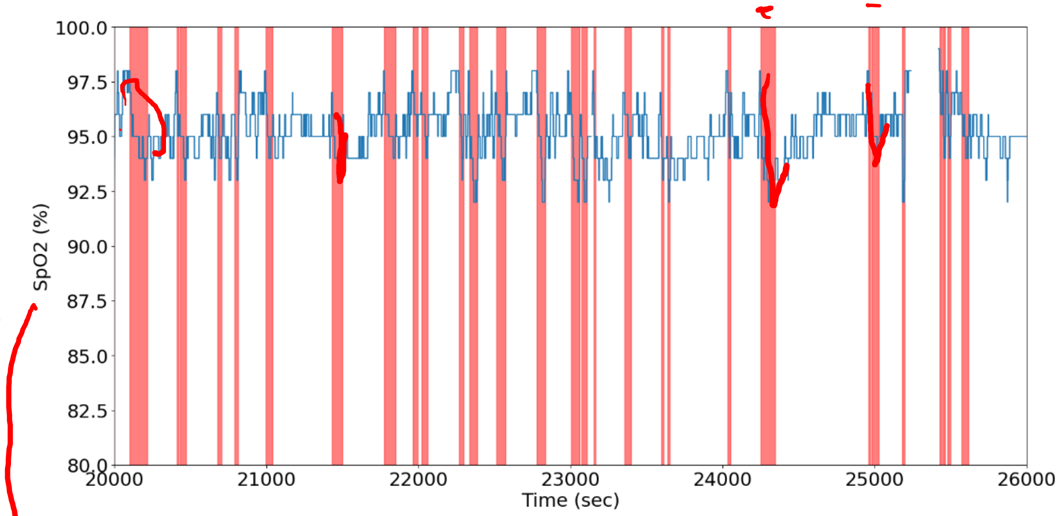
Examples: sleep apnea diagnosis

AIMLab.

OSA

- Detecting obstructive sleep apnea from biomarkers, demographics and anthropometric information.



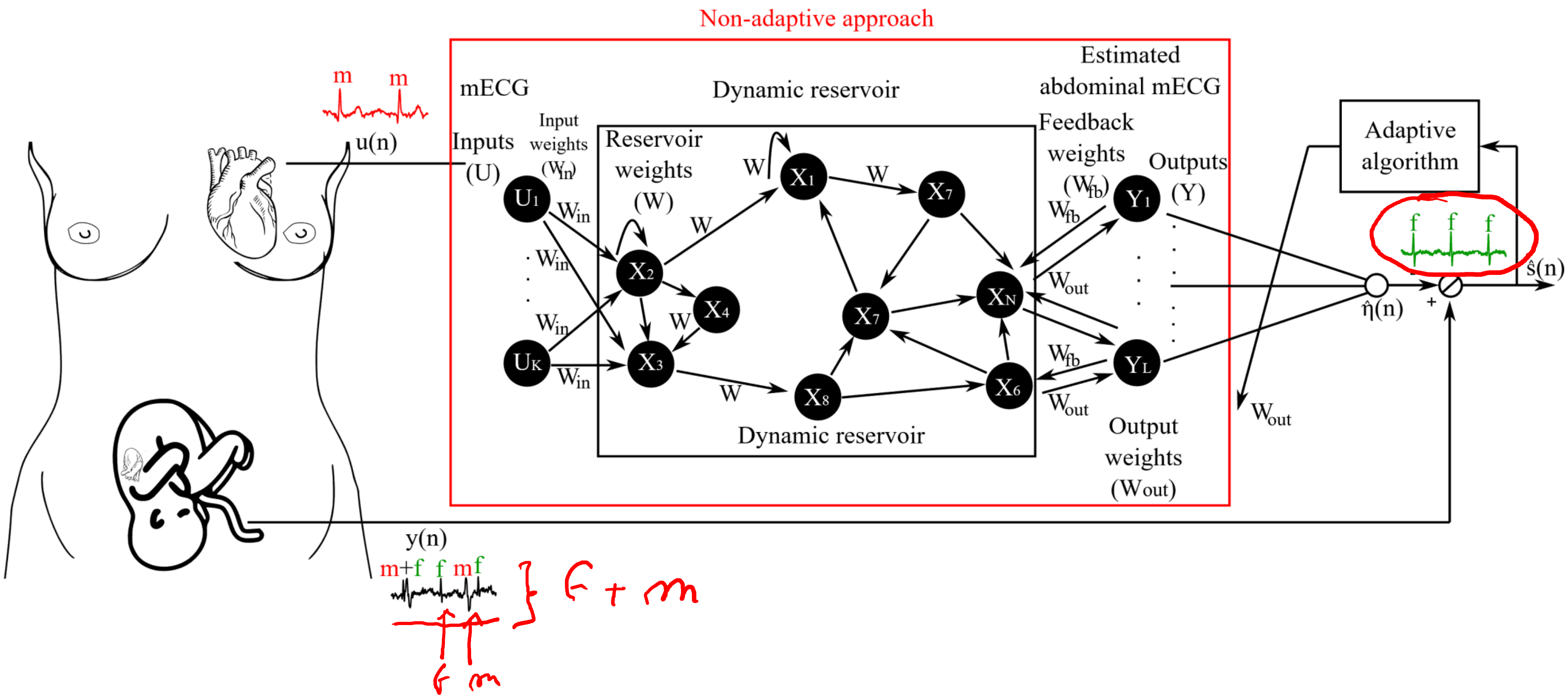


$$\frac{[HbO_2]}{[Hb]+[HbO_2]}$$

Behar, Joachim A., et al. "Feasibility of Single Channel Oximetry for Mass Screening of Obstructive Sleep Apnea." EClinicalMedicine (2019).

OBGY: fetal electrocardiography

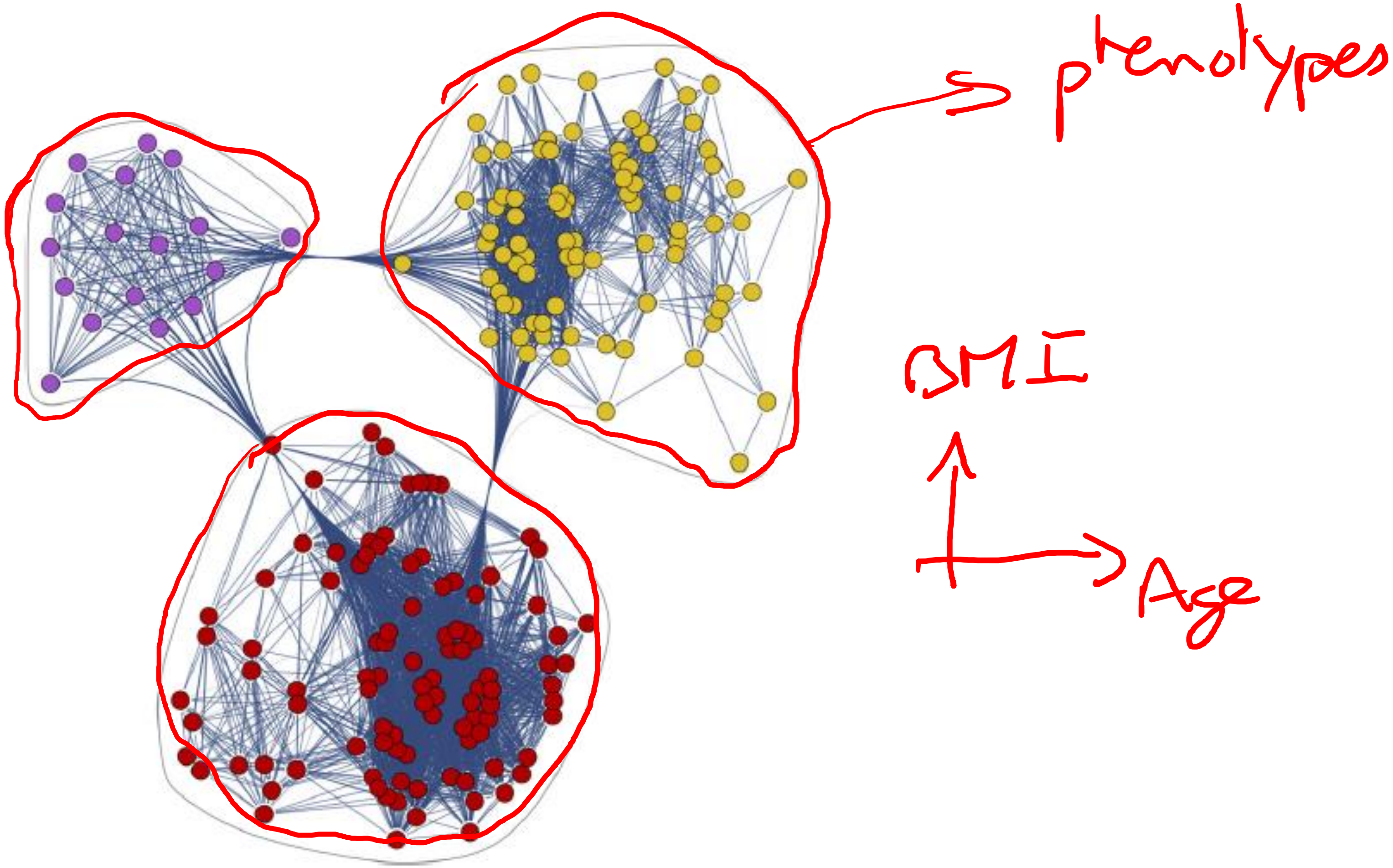
- Non-invasive fetal ECG diagnosis.



Kahankova R, Martinek R, Jaros R, Behbehani K, Matonia A, Jezewski M, Behar JA. A review of signal processing techniques for non-invasive fetal electrocardiography. IEEE reviews in biomedical engineering. 2019 Aug 29;13:51-73.

Unsupervised learning for disease phenotyping

AIMLab.



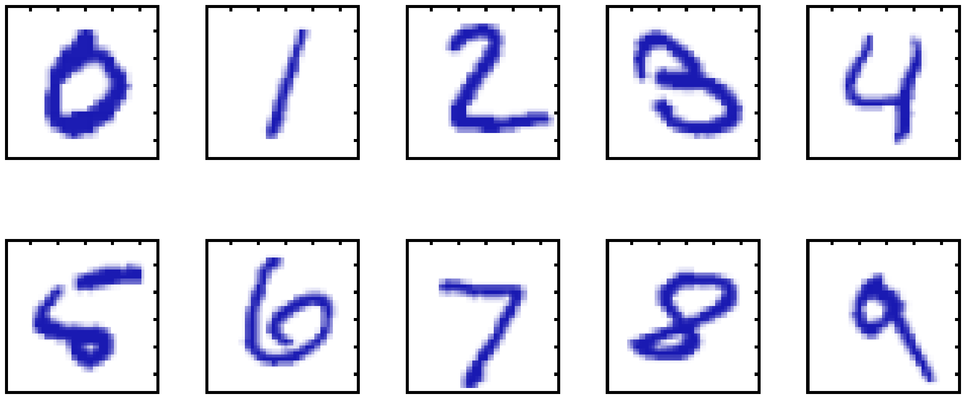
Lacedonia, Donato, et al. "Characterization of obstructive sleep apnea–hypopnea syndrome (OSA) population by means of cluster analysis." Journal of sleep research 25.6 (2016): 724-730.

Important concepts in ML

Vocabulary

0 $x^{(1)}$, 0 $x^{(2)}$, 1 $x^{(3)}$, 1 ... 1

- Consider a set of m **training examples:**
  - $\left\{x_{train}^{(1)}, x_{train}^{(2)}, \dots, x_{train}^{(m)}\right\},$
- Consider the corresponding **target labels**:
  - $\left\{y_{train}^{(1)}, y_{train}^{(2)}, \dots, y_{train}^{(m)}\right\}$



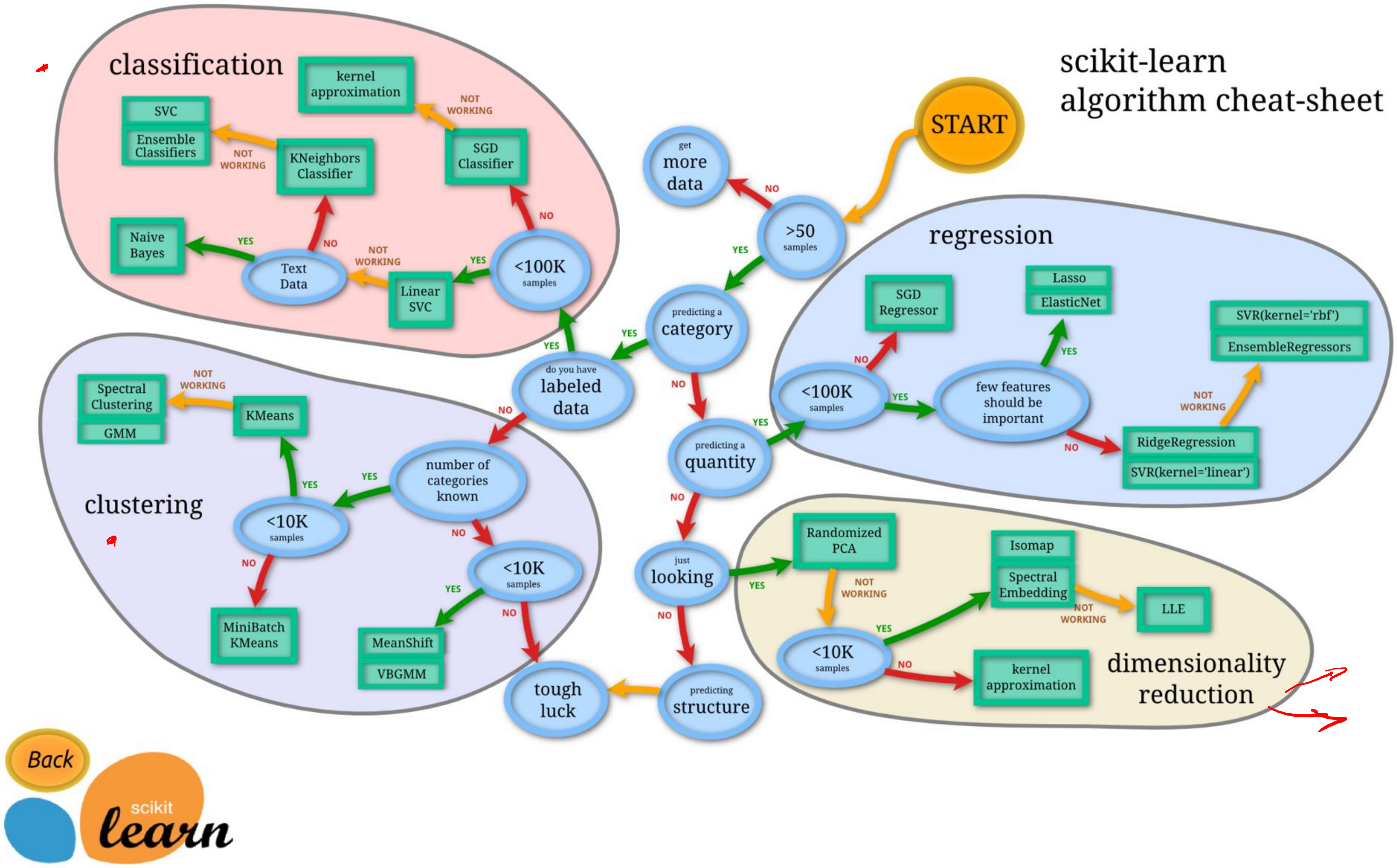
- We seek the **target function** f so that:
  - $f\colon\ x^{(k)} \rightarrow y^{(k)}$
- This function is learned during the learning phase by using the **training set**. In a second step, we seek to evaluate it on an unseen set of examples that we will call the **test set**:
  - $\left\{x_{test}^{(1)}, x_{test}^{(2)}, \dots, x_{test}^{(p)}\right\}, \left\{y_{test}^{(1)}, y_{test}^{(2)}, \dots, y_{test}^{(p)}\right\}$ } ML evaluation
  - The ability to accurately classify unseen examples is known as **generalization**.

Categories of ML

AIMLab.

- **Supervised Learning**: the training data comprises examples of the input vectors along with their corresponding target vector.
  - Classification: output is one or finite number of discrete categories.
  - Regression: output is one or more continuous variables.
- **Unsupervised Learning**: input vectors but no target values.
  - Clustering: to discover groups of similar examples within the dataset.
  - Density estimation: to estimate the distribution of the data within the input space.
  - Visualization: project data from high-dimensional space down to two or three dim.
- **Reinforcement Learning**: finding suitable actions to take in a given situation in order to maximize a reward.

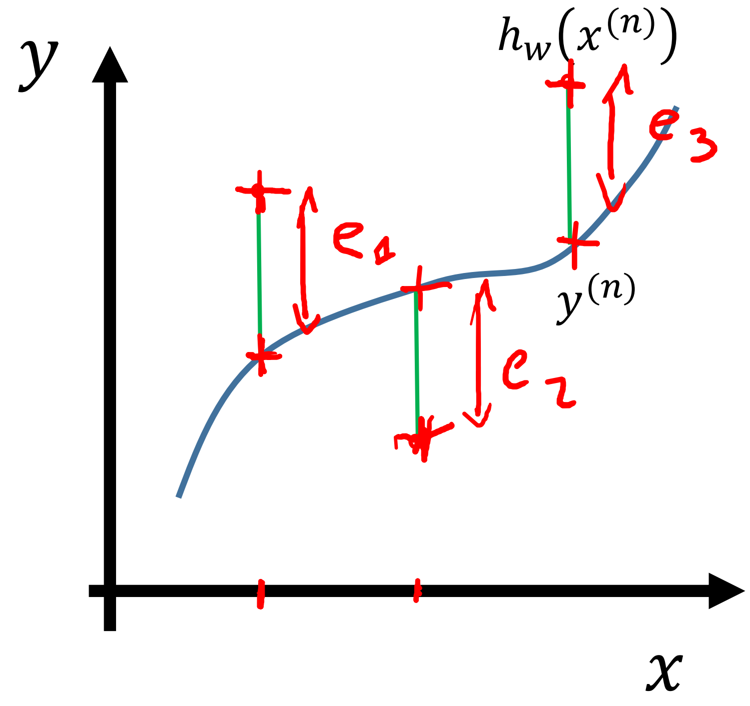
AIMLab.



scikit-learn
algorithm cheat-sheet
START
>50 samples
get more data
predicting a category
do you have labeled data
predicting a quantity
just looking
predicting structure
tough luck
classification
kernel approximation
SVC
Ensemble Classifiers
KNeighbors Classifier
SGD Classifier
Naive Bayes
Text Data
Linear SVC
<100K samples
clustering
Spectral Clustering
GMM
KMeans
number of categories known
<10K samples
MiniBatch KMeans
MeanShift
VBGMM
regression
SGD Regressor
Lasso
ElasticNet
SVR(kernel='rbf')
EnsembleRegressors
few features should be important
RidgeRegression
SVR(kernel='linear')
dimensionality reduction
Randomized PCA
Isomap
Spectral Embedding
LLE
kernel approximation
YES
NO
NOT WORKING
Back
scikit learn

Other important concepts

- We will use the curve fitting toy example.
- Consider a set of training examples and their associated target:
  - $\left\{x_{train}^{(1)}, x_{train}^{(2)}, \dots, x_{train}^{(m)}\right\}$,
  - $\left\{y_{train}^{(1)}, y_{train}^{(2)}, \dots, y_{train}^{(m)}\right\}$



- We will consider a polynomial **hypothesis function**:
  - $h_w(x) = w_0 + w_1 x + w_2 x^2 + \dots + w_M x^M$
  - M is the order of the polynomial.

$\{w_i\}_{i \in [1,M]}$

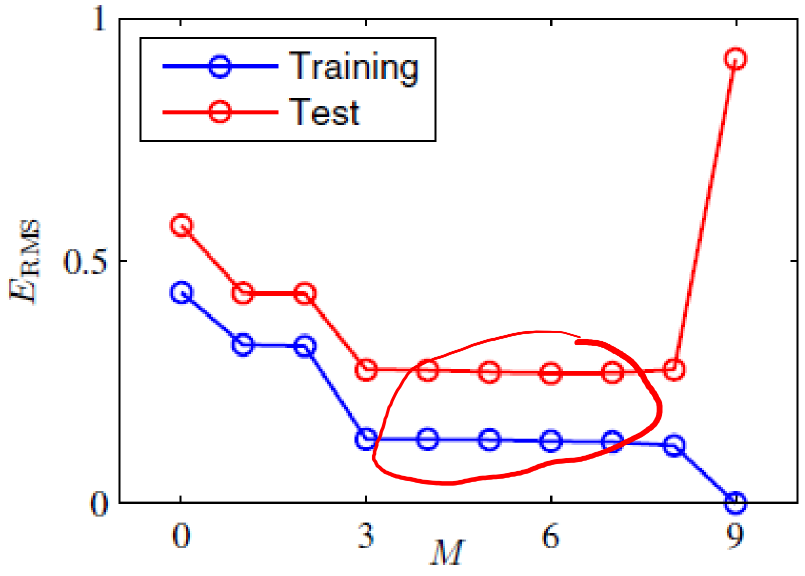
- The value of the coefficients will be determined by fitting the polynomial to the training data. This will be achieved by minimizing an **error function** such as the mean square error:
  - $J(w) = \frac{1}{m}\sum_{n=1}^{m}\{h_w\left(x^{(n)}\right) - y^{(n)}\}^2$

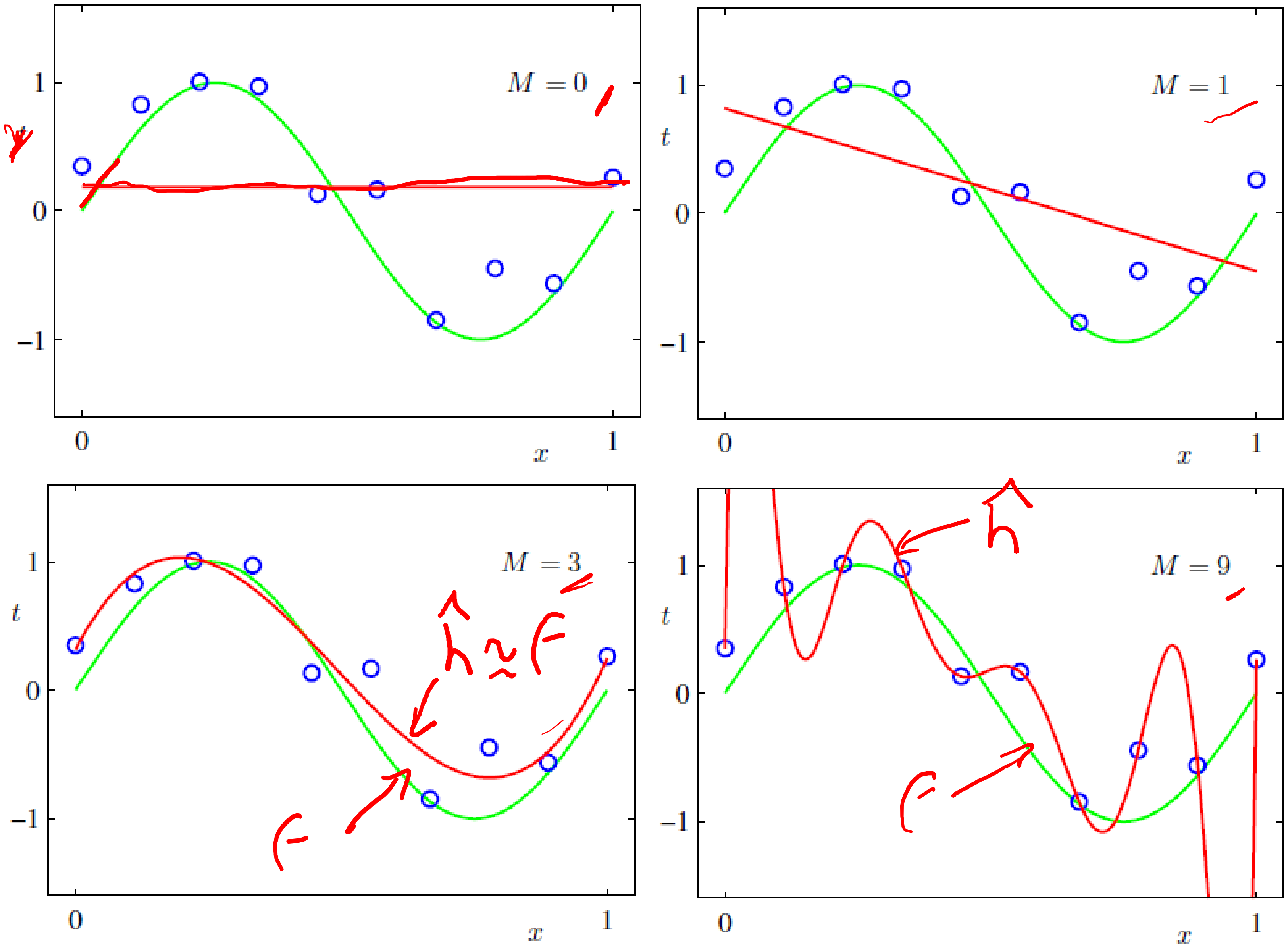
$\rightarrow \min_w \left(J(w)\right)$

Model selection

AIMLab.

- Under fitting (high bias),
- Overfitting (high variance).





Images: Bishop, Christopher M. Pattern recognition and machine learning. springer, 2006.

Machine Learning

Unknown
Target Function: $\boldsymbol{f: x \rightarrow y}$

Training Examples
$\{x^{(1)}, x^{(2)}, \dots, x^{(m)}\}$
$\{y^{(1)}, y^{(2)}, \dots, y^{(m)}\}$

Hypothesis Set
$H = \{h\}$

Learning Algorithm

Final Hypothesis
$\hat{h} \approx f$

$x^{(i)}$
i = example nb

$\hat{h} \approx F$

References

[1] Deep Learning Course notes. Prof. Gilles Louppe. Universite de Liege. Spring 2019.

[2] Introduction to Machine Learning. Doron Shaked, Sagi Schein, Omer Barkol. BME course notes 2018.

[3] Pattern Recognition and Machine Learning. Springer 2006. Christopher M. Bishop.