Hes·so
Haute Ecole Spécialisée
de Suisse occidentale
Fachhochschule Westschweiz
University of Applied Sciences
Western Switzerland

MSE | MASTER OF SCIENCE IN ENGINEERING

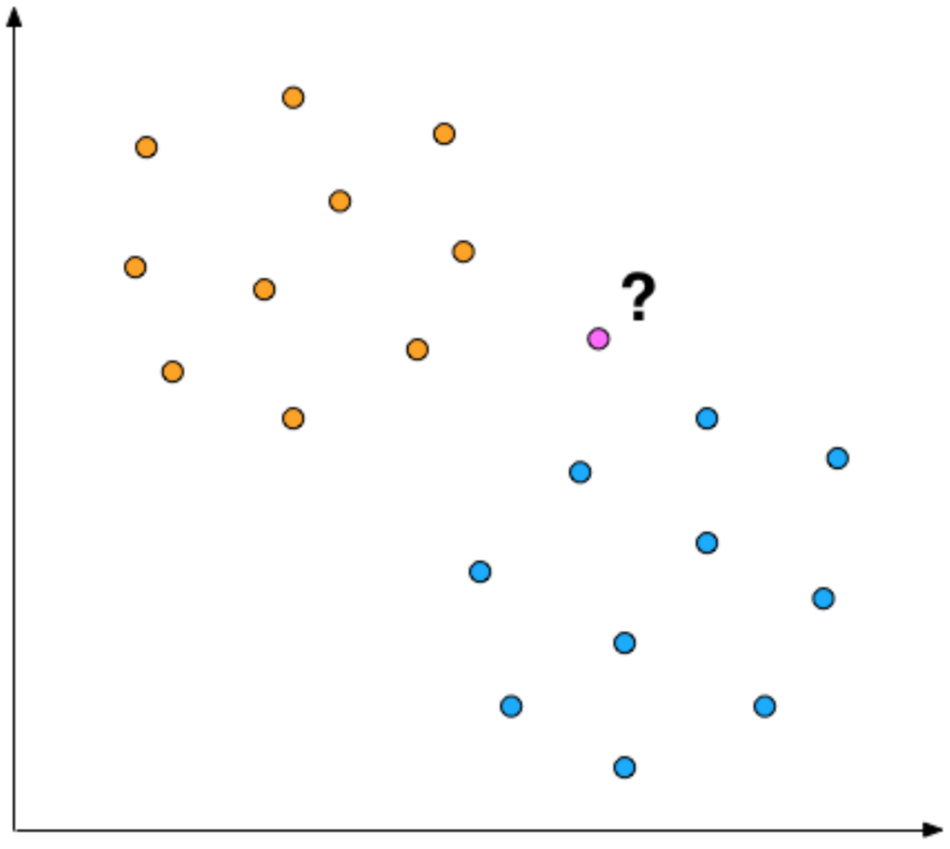
Multimodal Processing, Recognition and Interaction

I. The classification problem
II. Introduction to the Hidden Markov Models & Time Series

Stefano Carrino, Elena Mugellini, Omar Abou Khaled

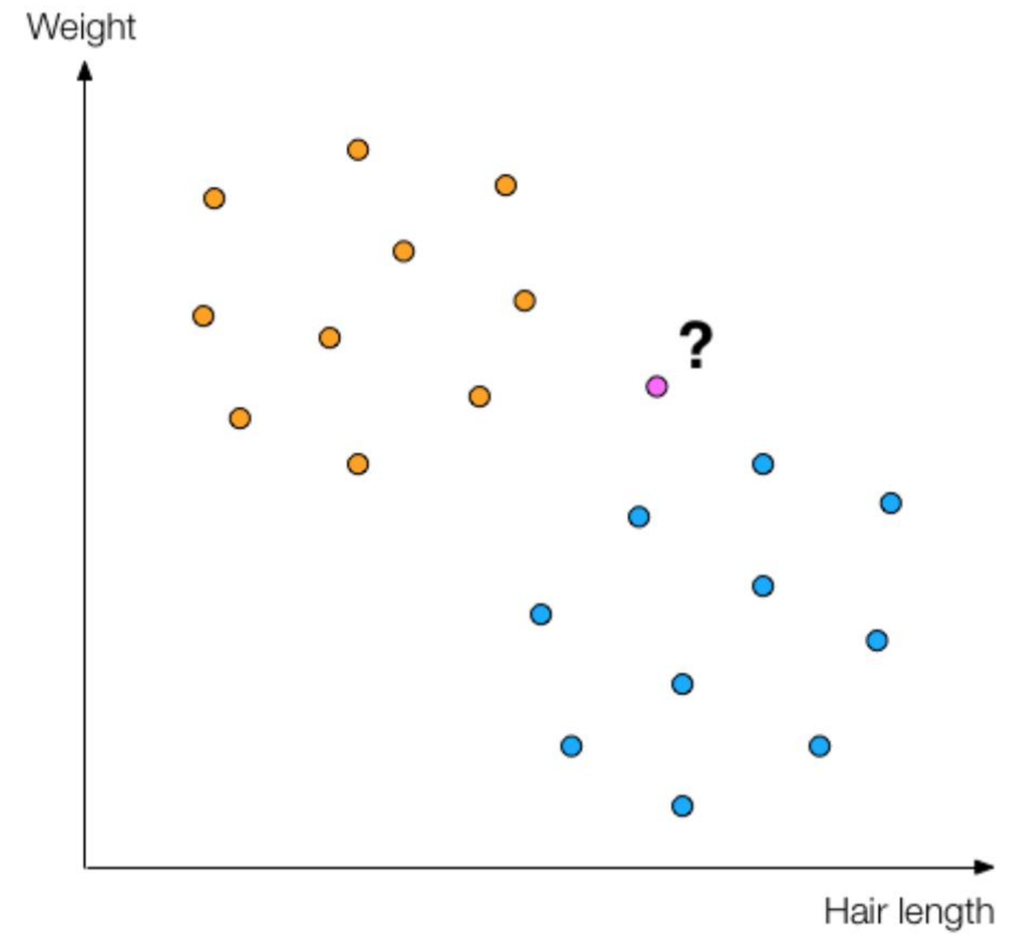
Classification problem

- Common task in Machine Learning
- Given data points belonging to several **classes** (here 2), the goal is to predict (decide) in which class a new point will be

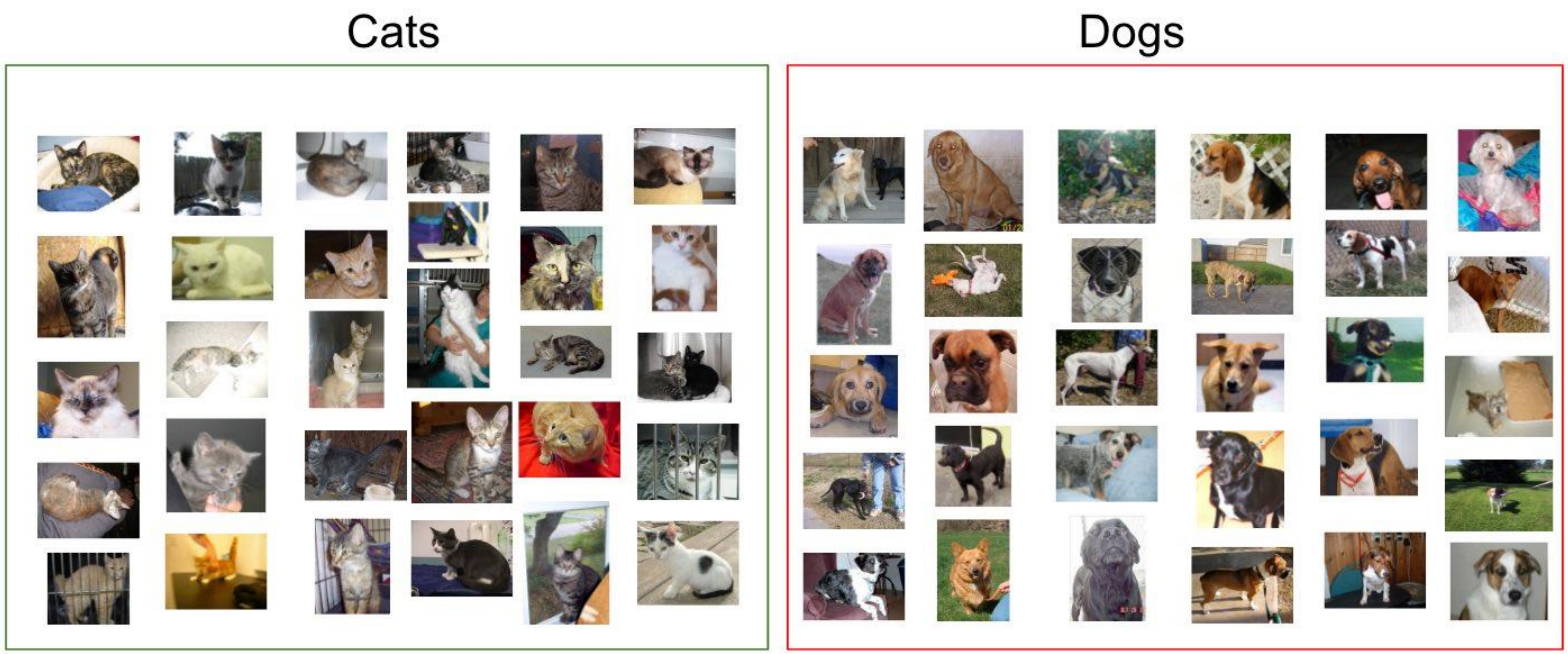


Classification problem

- Example - 2 classes problem:
  - Women (●)
  - Men (●)
- 2 features:
  - Hair length (axe X)
  - Weight (axe Y)
- What about the new sample (●)? Is it a woman or a man?



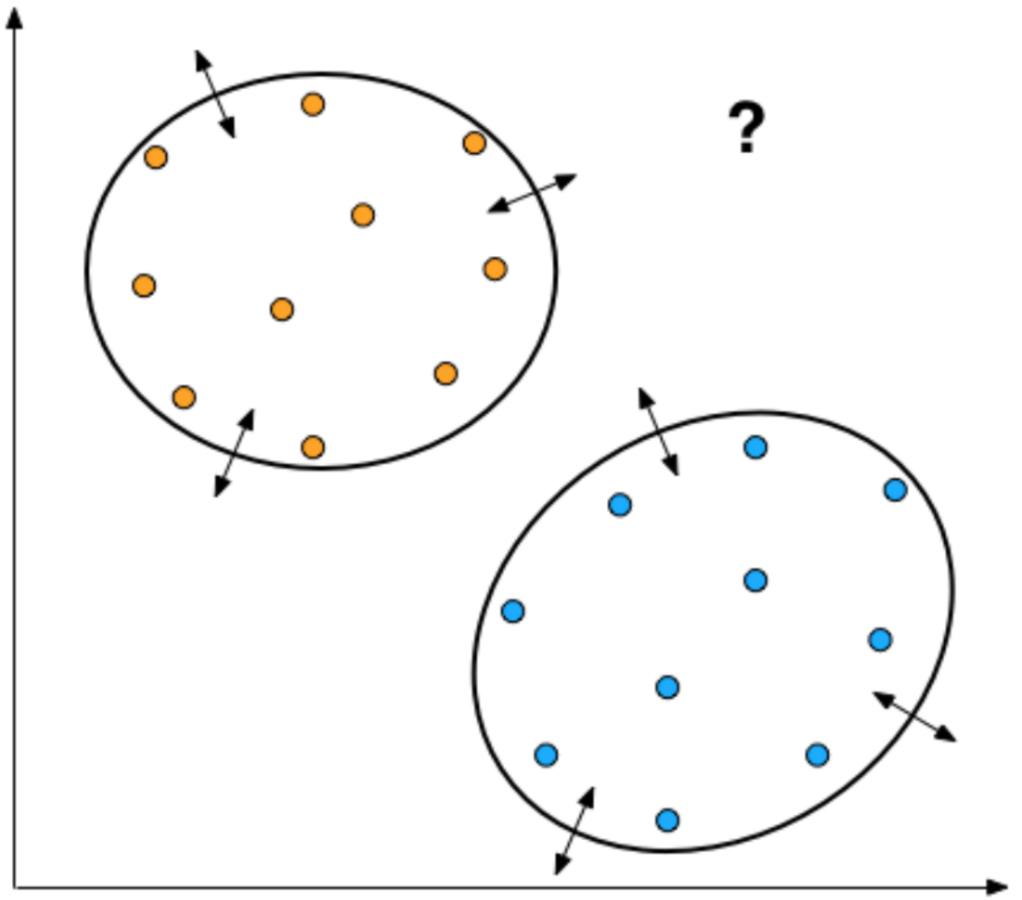
Classification problem



Sample of cats & dogs images from Kaggle Dataset

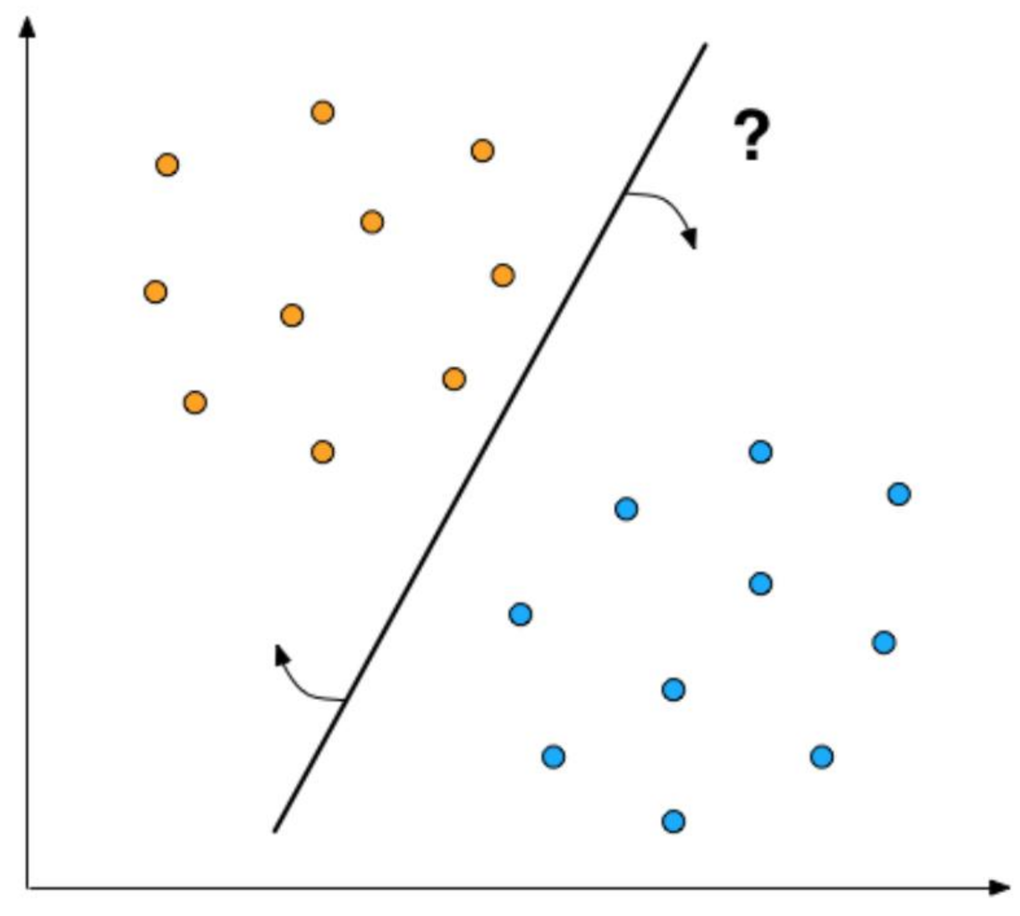
Classification problem – Generative approach

- Also called **Class modeling approach**
- Some algorithms try to **model classes**, i.e. to group samples by classes
- E.g. Gaussian Mixture Models (**GMM**)
- Which are the **best models** for a class?

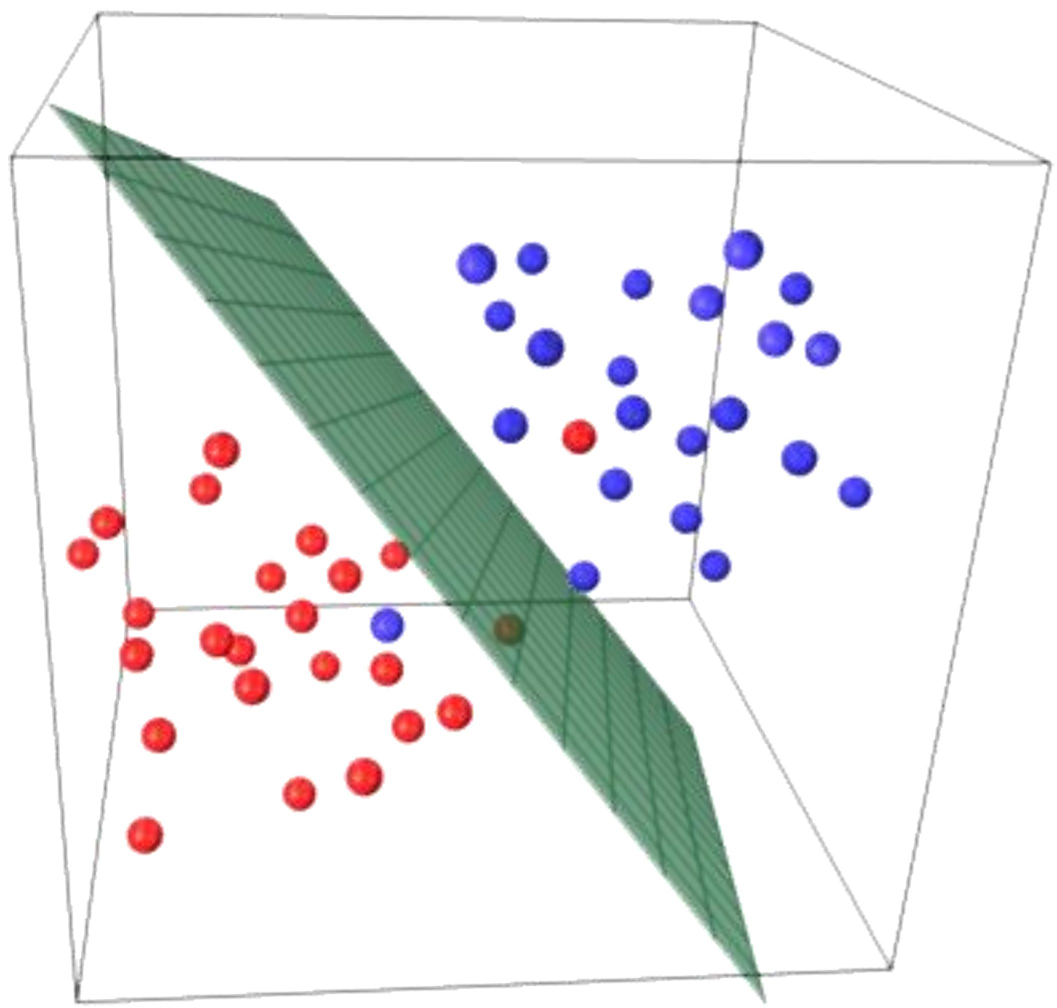


Classification problem – Discriminative approach

- Some algorithms try to **discriminate classes**, i.e. to separate samples by classes
- E.g. Support Vector Machines (**SVM**)
- Which is **the best separating hyperplane**?

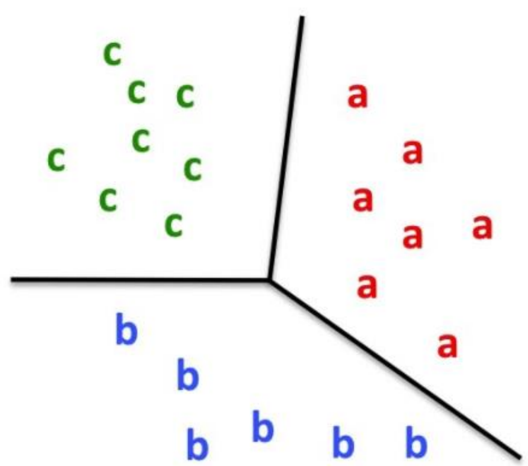


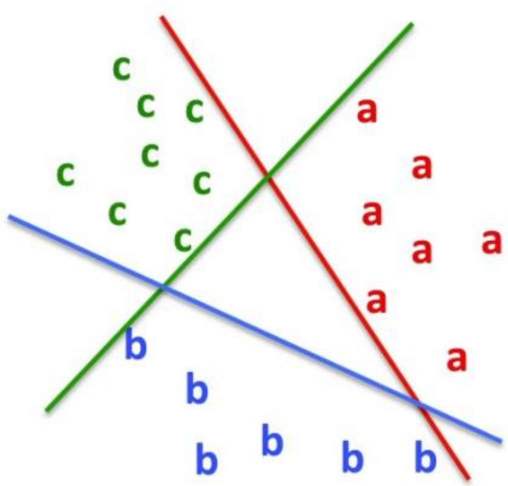
Hyperplane



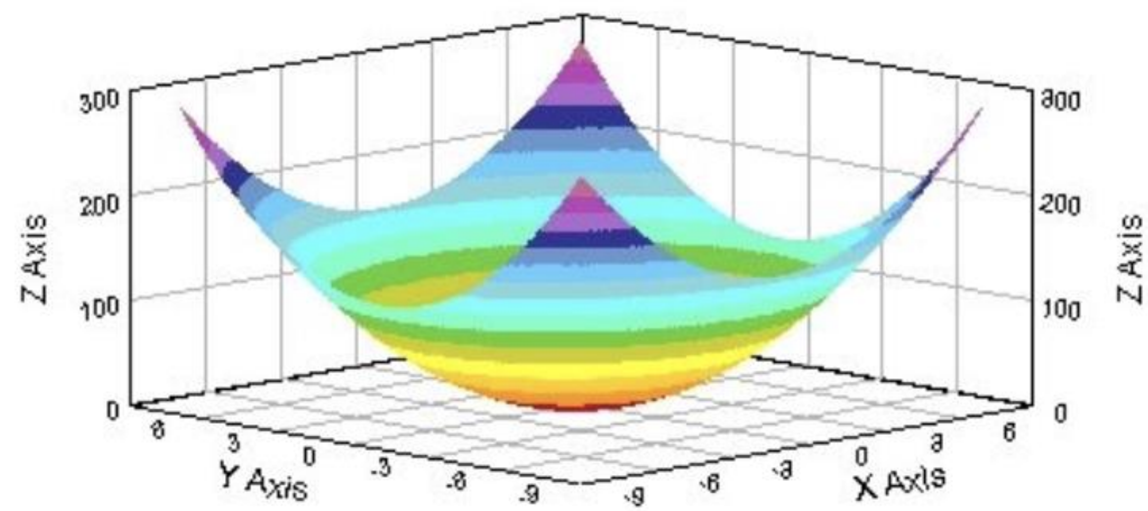
**More about this topic when we will talk about SVM*

Multi-class & Multi-dimensional (or Multivariate) Classification problems





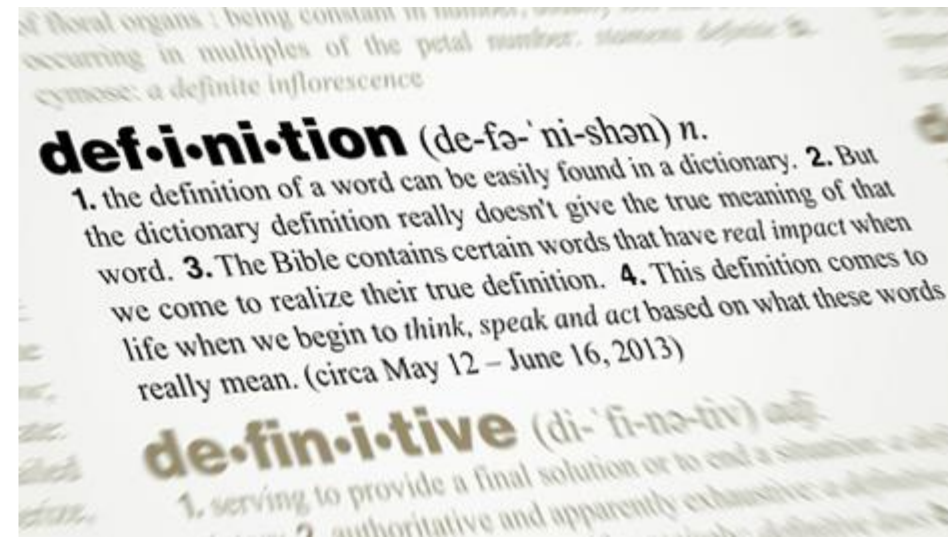
*Classes can be more than 2. Think of **object recognition** or automated translation software.*

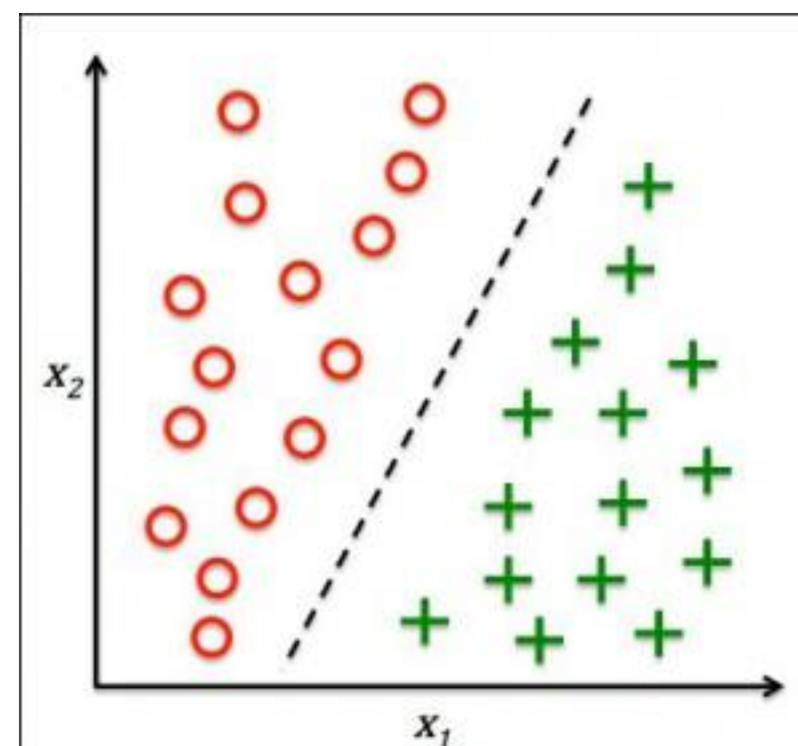


Difficult to draw more than 3 dimension.
Learning problems with 2, 3, 10’000+ dimensions are common.
***Computer vision**: 100 x 100 picture - every pixel is a feature.. times 3 dimensions! (RGB)*

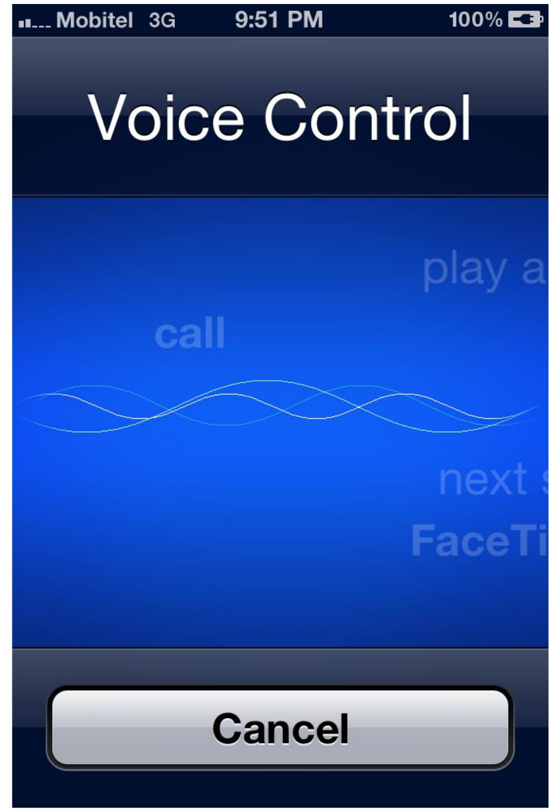
Classification definition

In <u>supervised</u> machine learning, **classification** is the problem of identifying to which of a **set of categories** (sub-populations) a new observation belongs, on the basis of a **training set** of data containing observations (or instances) whose category membership is known.

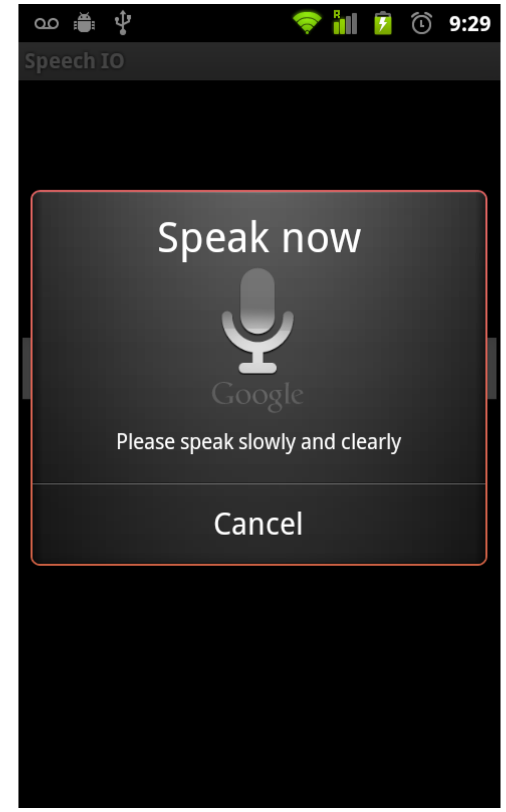




Siri



Voice Control



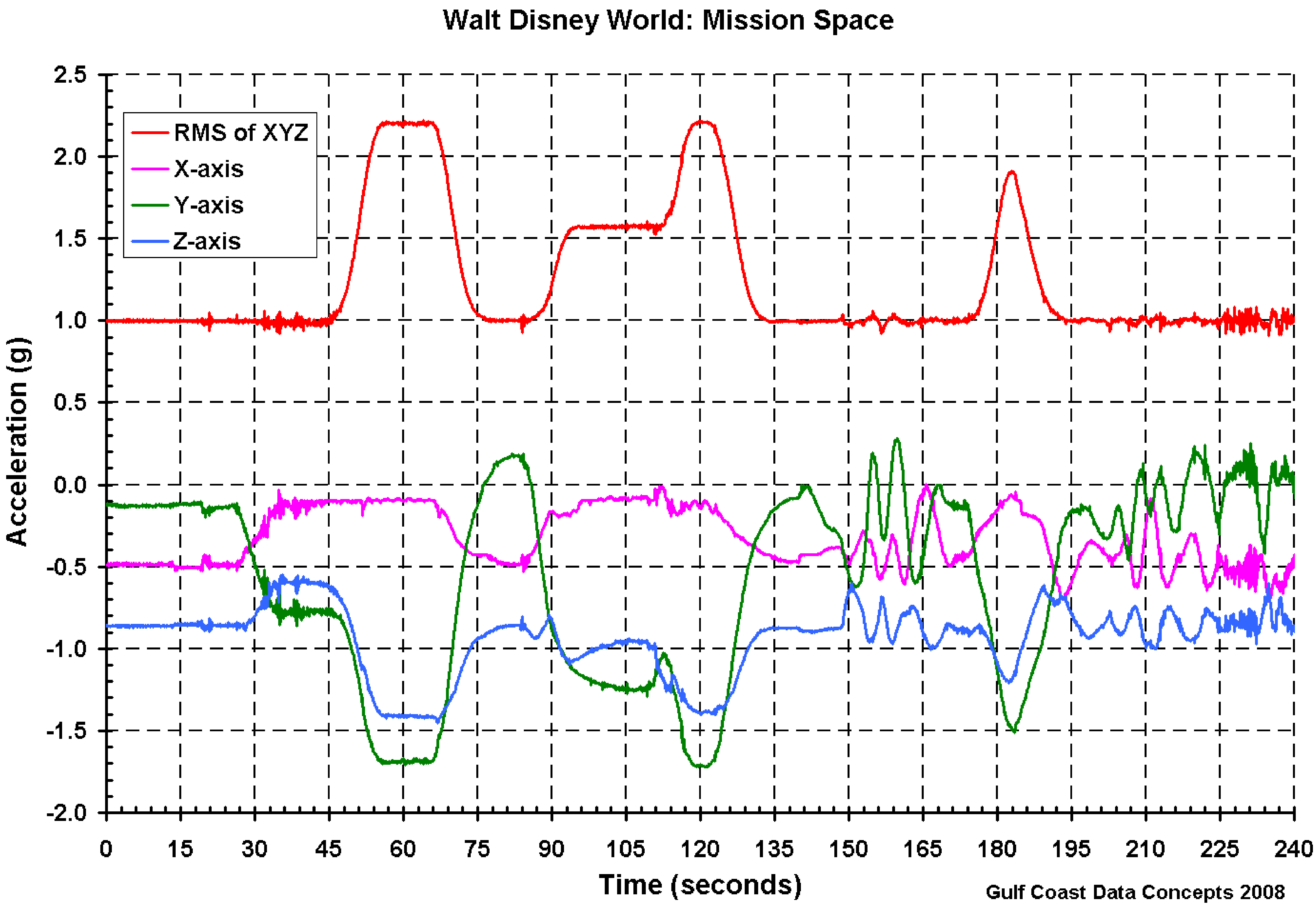
Motivation: Time Series

Use Cases: Gesture recognition & Speech processing

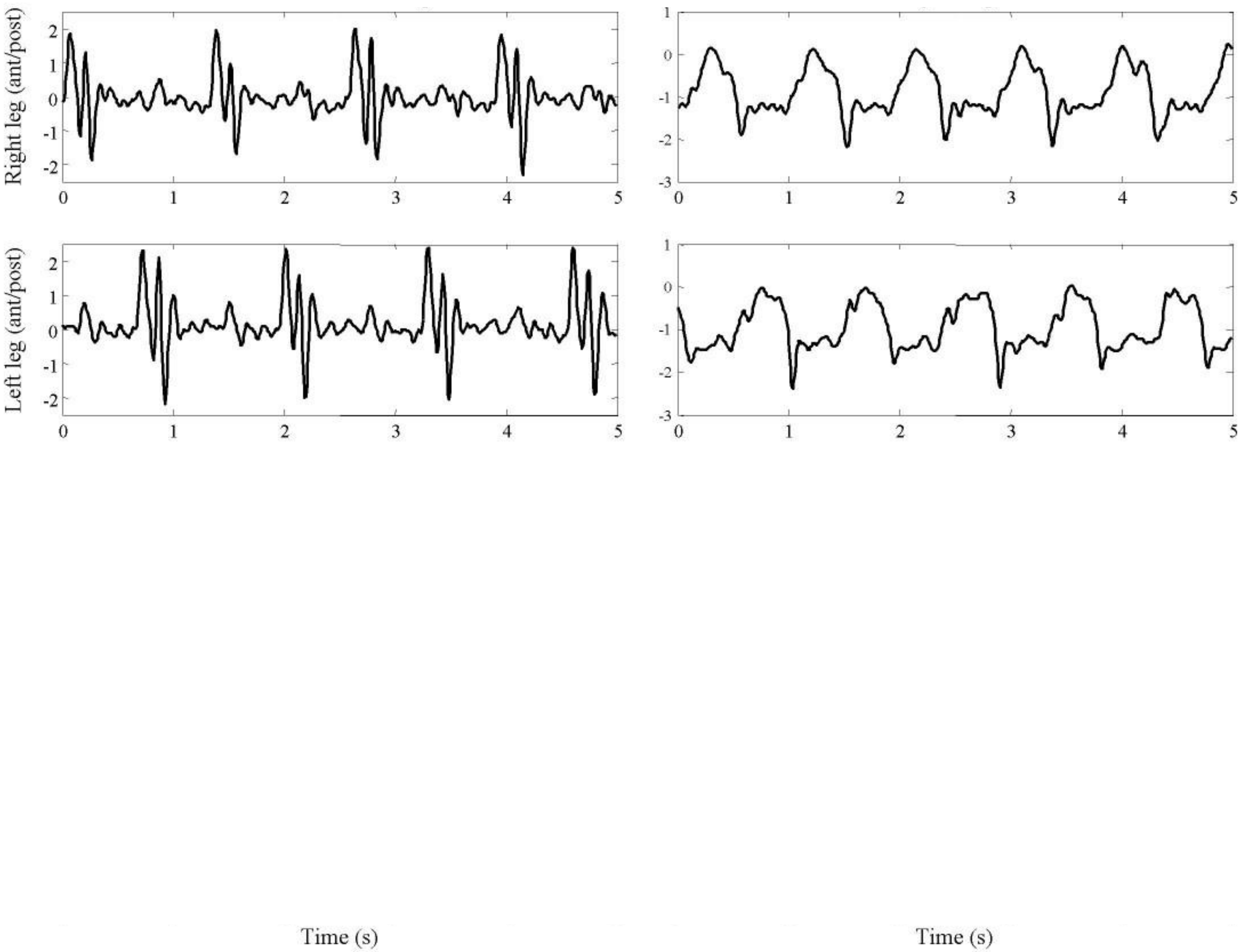
INTRODUCTION TO HIDDEN MARKOV MODELS (HMMS)

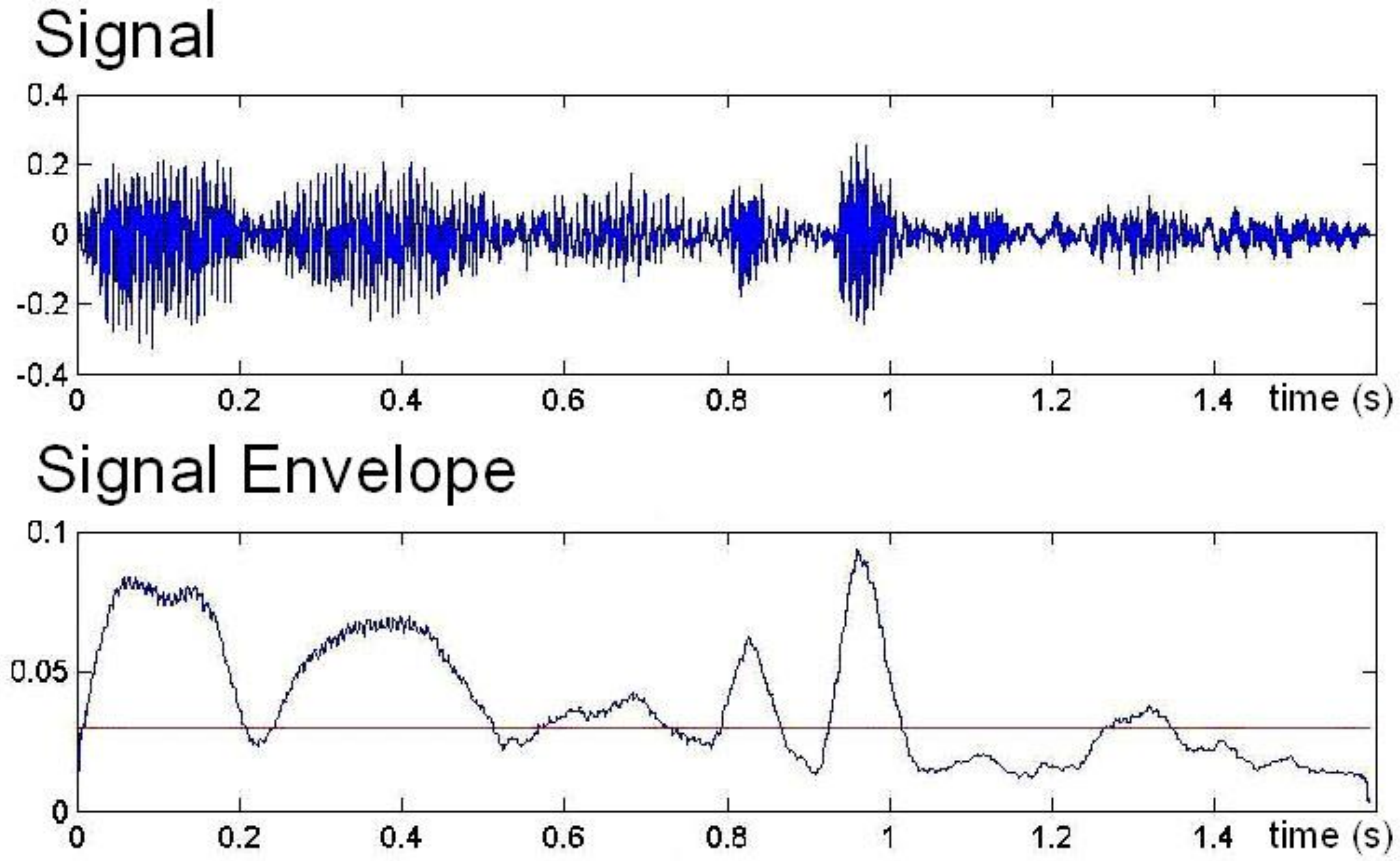
Motivations and Challenges

TIME SERIES

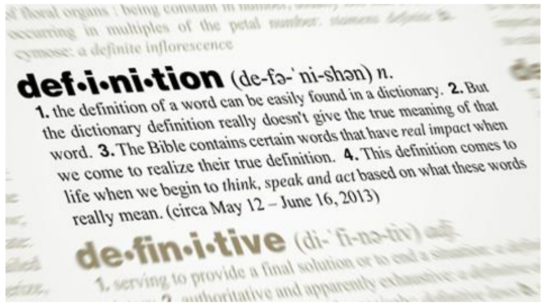


Source: http://www.gcdataconcepts.com/wdwxlr8r.html


Right leg (ant/post)
Left leg (ant/post)
Time (s)
Time (s)


Signal
0.4
0.2
0
-0.2
-0.4
0
0.2
0.4
0.6
0.8
1
1.2
1.4
time (s)
Signal Envelope
0.1
0.05
0
0
0.2
0.4
0.6
0.8
1
1.2
1.4
time (s)

Time Series – Definition



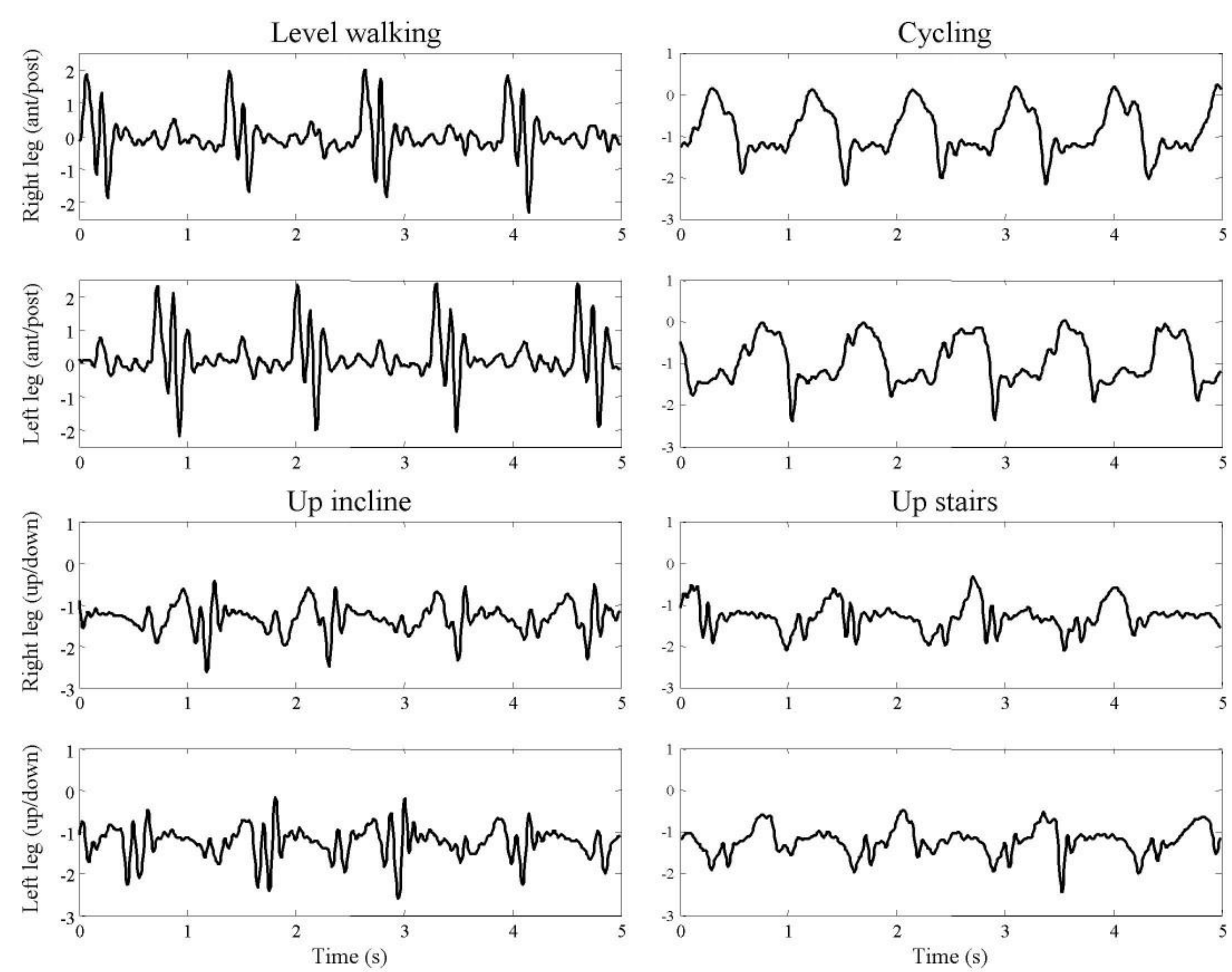
Time Series: *the sequence of observations x_t (with $t \in T$) of a variable x at different instants is called* ***time series****. Usually, T is countable, so that t = 1,…, T.*

Séries temporelles : *la suite d'observations x_t (avec $t \in T$) d'une variable x à différents temps est appelée* ***série temporelle****. Habituellement, T est dénombrable, de sorte que t = 1,…, T.*

Time Series – Motivation

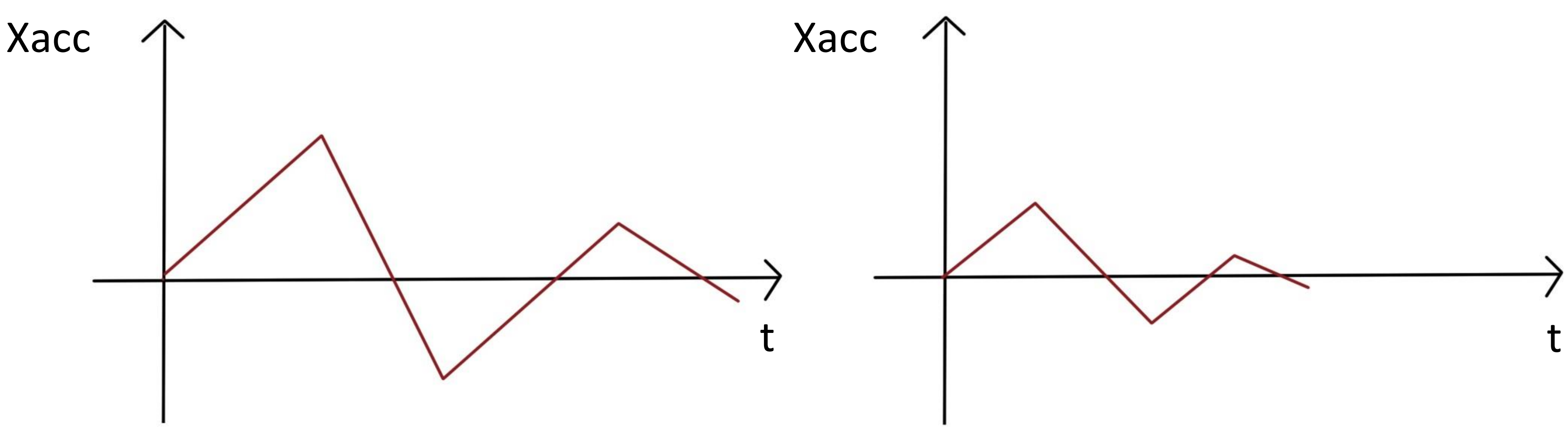
- Forecast (also regression)

- Classification



Time Series – Challenges (I)

- Variable length of a signal
  - E.g. gestures, word, sentences, etc.



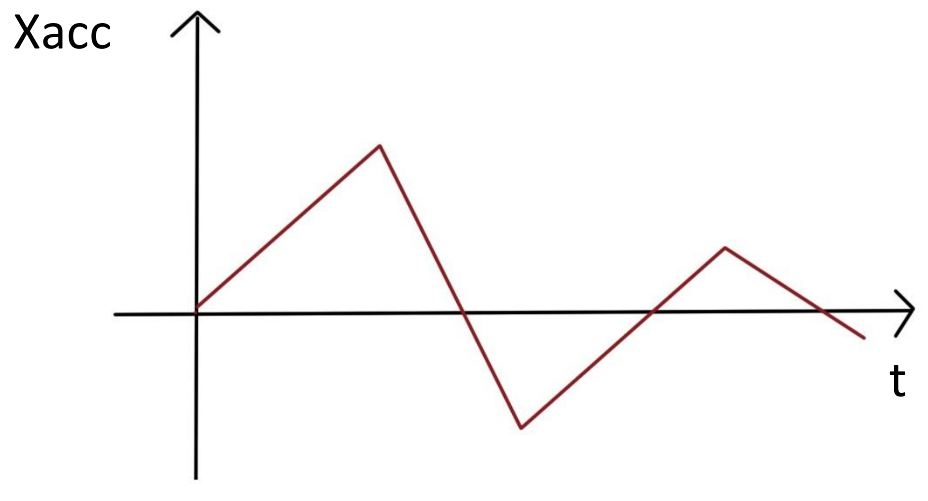
1s = 100 samples

0.7s = 70 samples

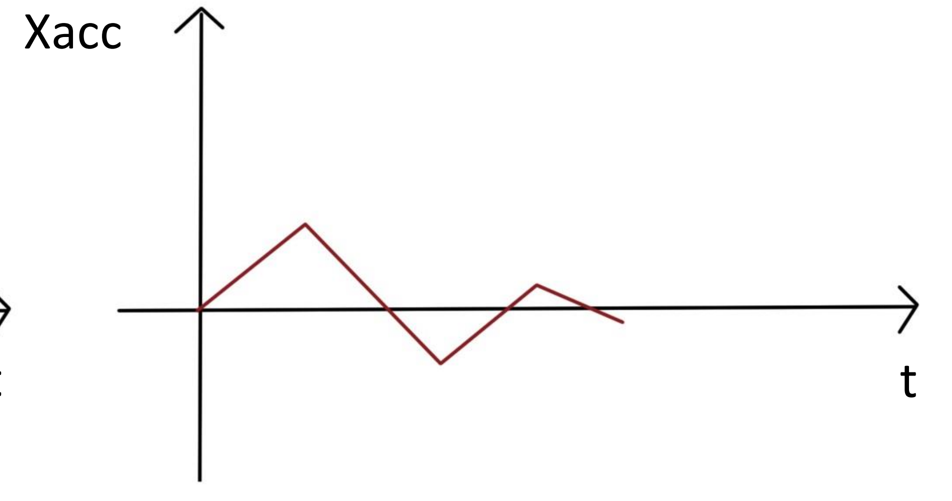
Time Series – Challenges (II)

- Signals are often complex
- Pre-processing is often needed to extract meaningful information

=> **Feature extraction**

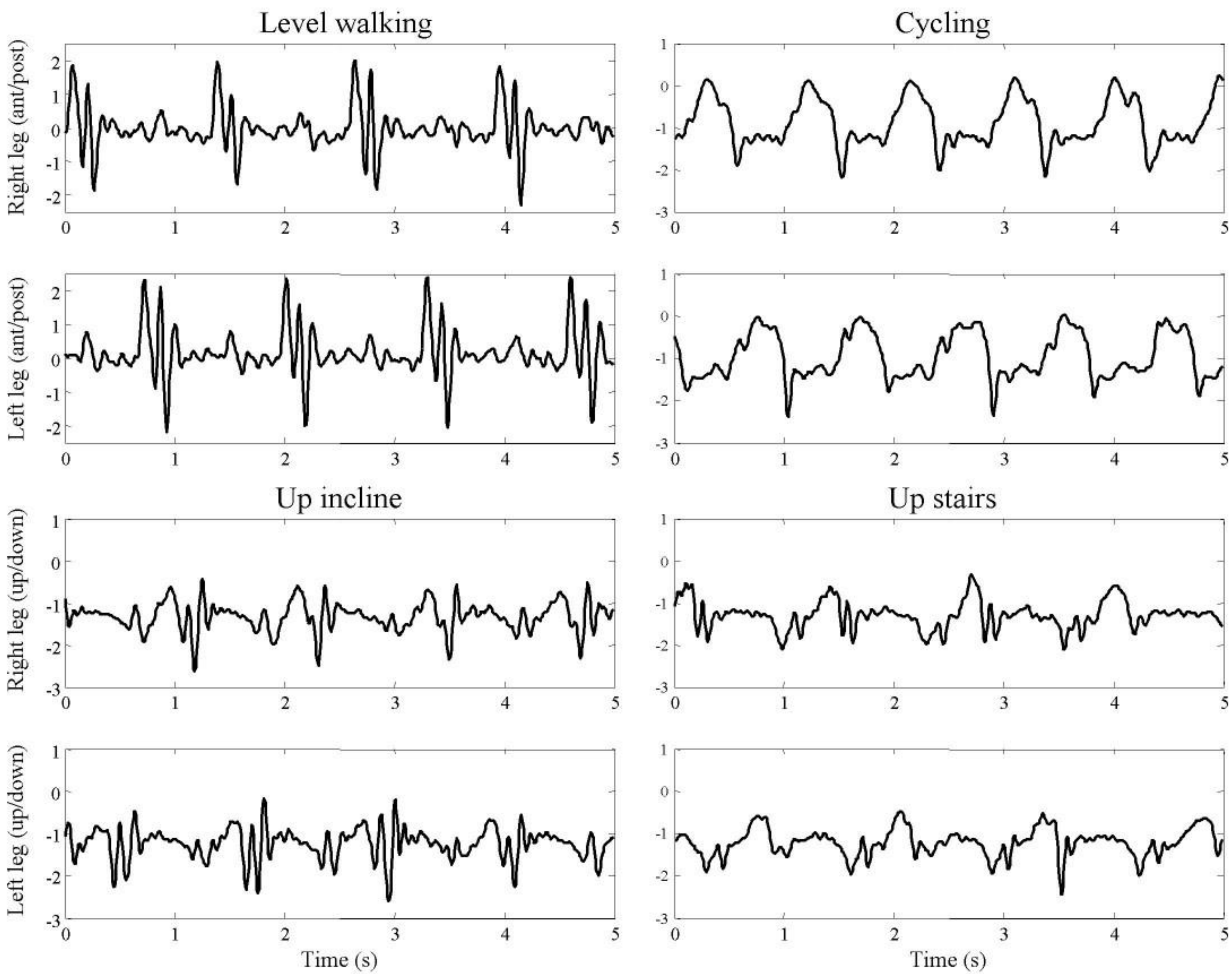


1s = 100 samples



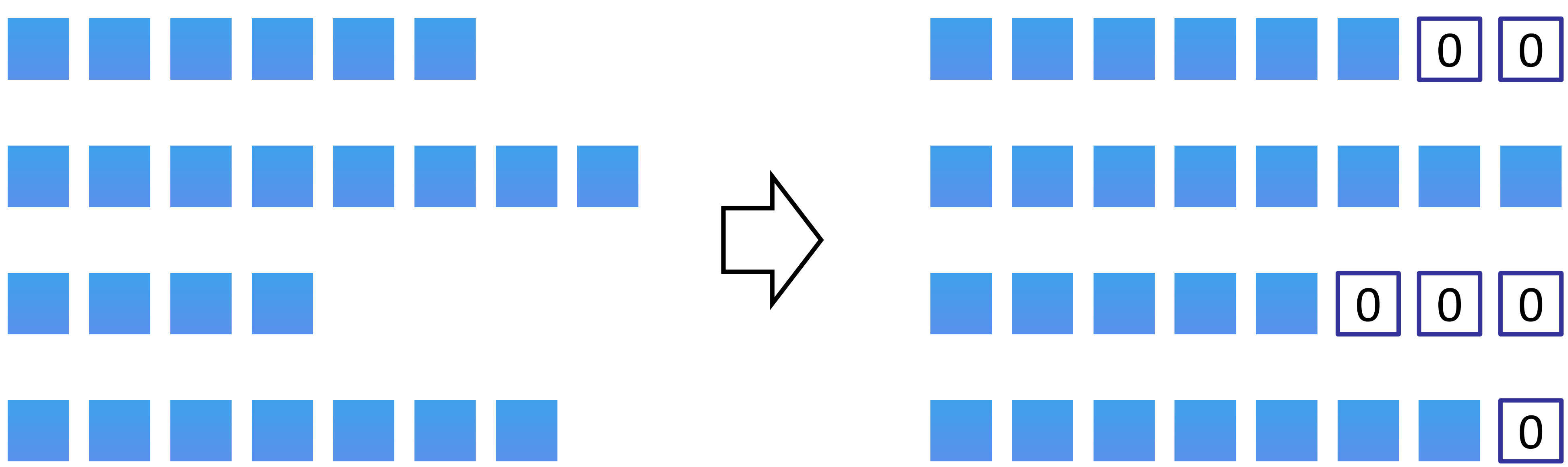
0.7s = 70 samples

Exercise:
Which features would **you** extract?



Time Series – Challenges (III)

- Solutions:
  - “Holistic” approaches
    - General characterization of a signal: max, min, mean, duration, etc.
  - Resampling
  - Padding



Example: “zero” padding

Time Series – Challenges (IV)

- Solutions:
  - Use of machine learning techniques that can directly deal with (can model) time series
    - HMM (Hidden Marcov Model), CRF (Conditional Random Fields), etc.

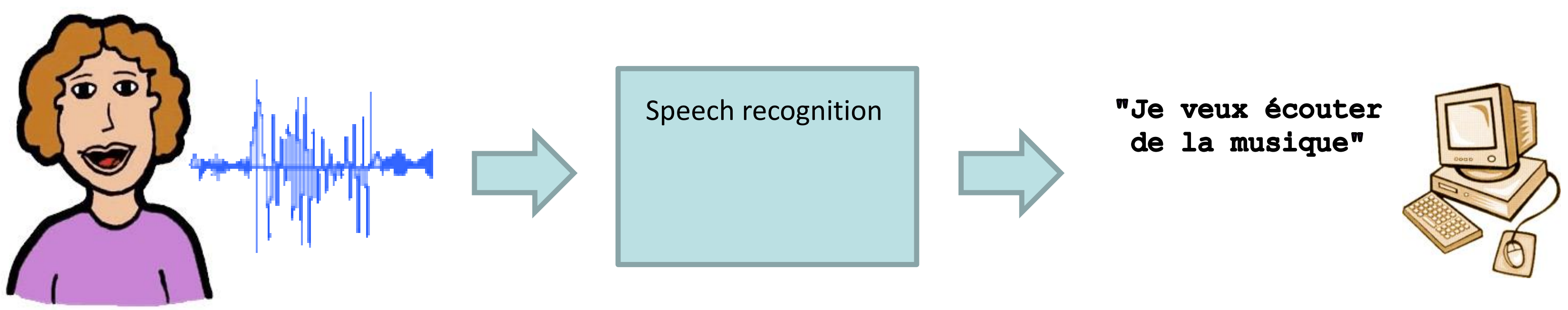
A specific case of time series: Speech

SPEECH PROCESSING

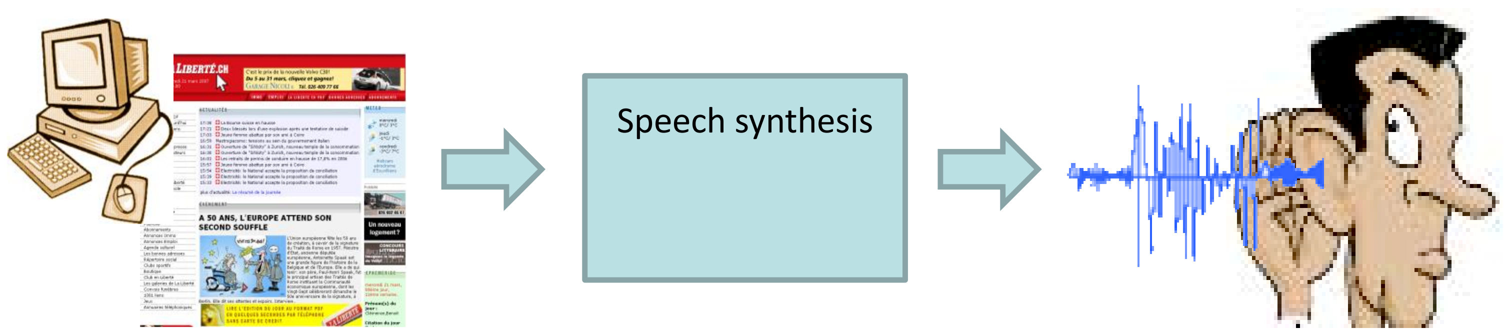
Voice-Based Interaction

- Why?
  - Speech a natural means of communication
  - Fast
  - Other kinds of interactions are not possible / convenient
    - Free-hand interaction (operating theater)

Voice-Based Interaction



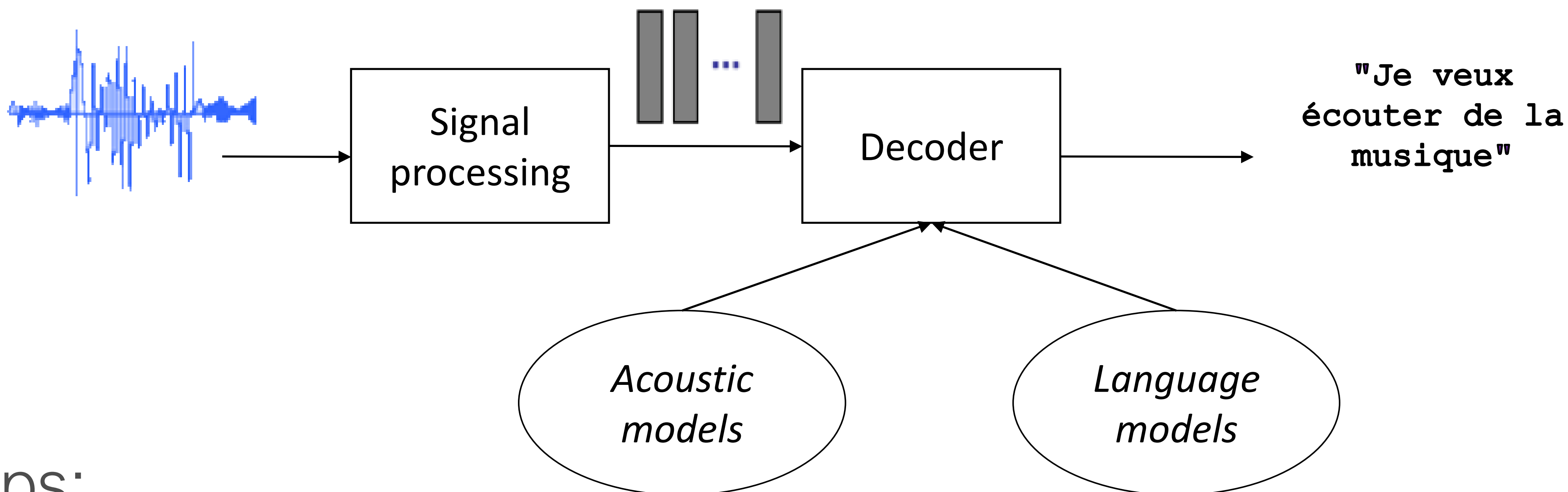
Speech recognition



Speech Synthesis

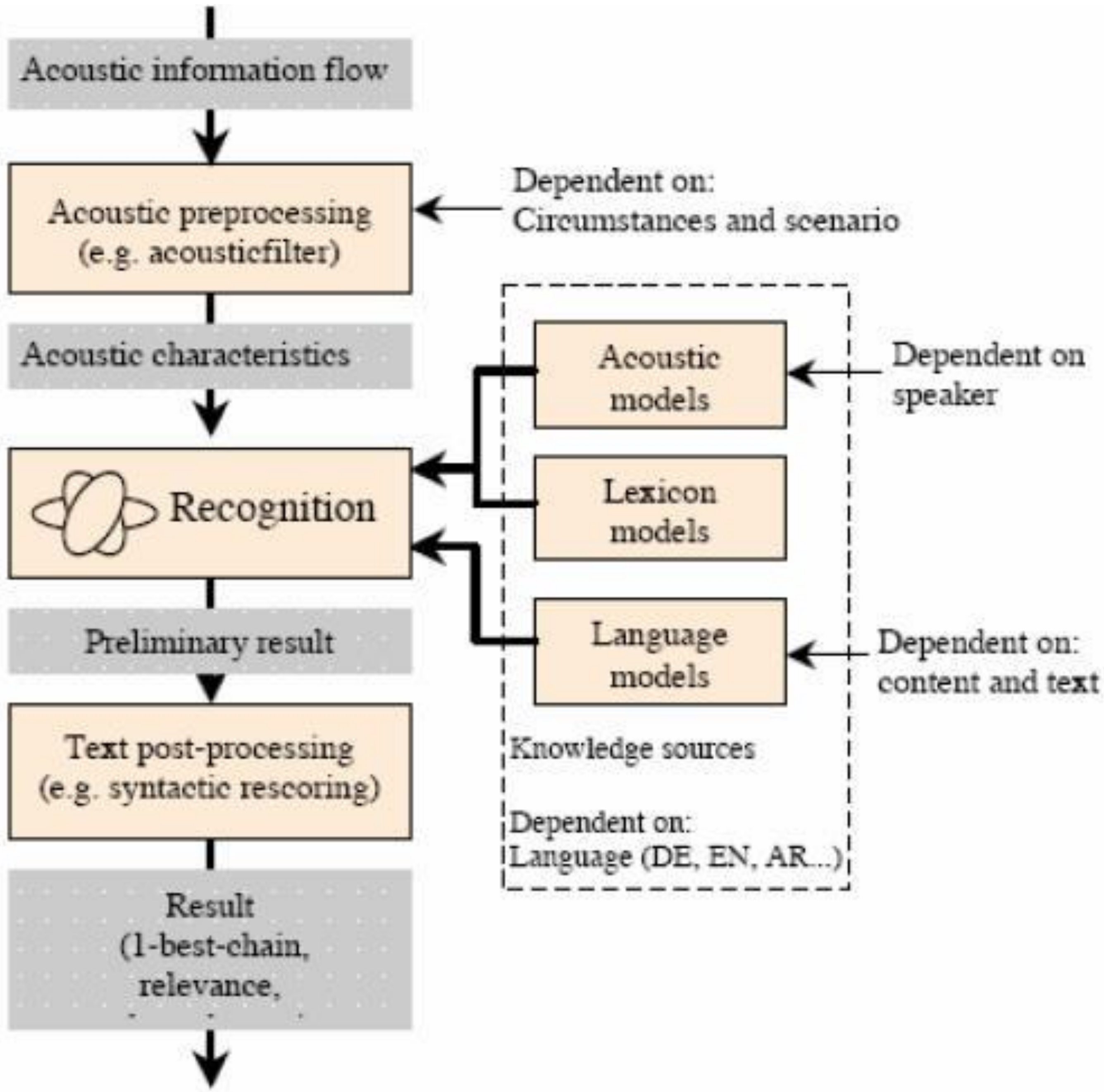
Dialogue

Speech Recognition



- Steps:
  - In a simple application based only on the acoustic model, the application will parse the pronounced word in **phonemes** who are the founding block of a word.
  - These phonemes are converted in digital “elements”.
  - Such a digital format, or *pattern,* is then classified using a machine learning approach

Speech Recognition

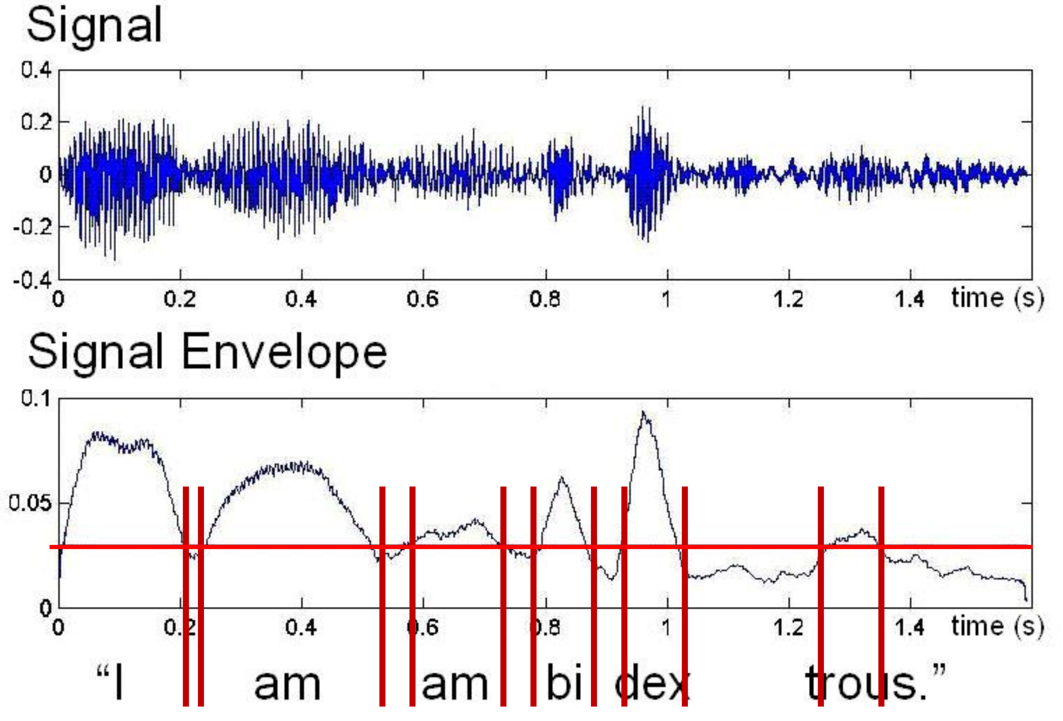


Speech Recognition – Parameters

- Speaker-Dependent Vs Speaker-Independent
  - S.D.:
    - Advantage: best results
    - The greater drawback is that such a system is focused on a single user and it has to be specifically trained on her/him.
  - S.I.:
    - Advantage: yes… general algorithm, valide for a wide population
    - Drawbacks: complexity, poorer performances (translation time, accuracy, …).
  - Speaker Adaptive

Speech Recognition – Parameters (II)

- Keyword spotting Vs Continuous speech
  - The recognition of isolated words is easier to model and realize, since the system knows the exact duration of each word
    - No need of *segmentation*



Speech Recognition – Parameters (III)

- Grammar
  - The grammar is used to define the valid words and also the underlying syntax
  - Grammar, consisting of a set of semantic and syntactic rules, is generally specified based on a set of conditions
- Vocabulary (small Vs. big)
  - Typically dependent on the application goal
  - Small vocabularies are easier to implement and recognize and require a smaller training set
  - Typical dictionary sizes: 10, 100, 1000, 10000 or 64000 words

Hes·so Haute Ecole Spécialisée de Suisse occidentale
Fachhochschule Westschweiz
University of Applied Sciences Western Switzerland

Speech Recognition – Parameters (IV)

Ex:

Dial three three two six five four

Dial nine zero four one oh nine

Phone Woodland

Call Steve Young

```
$digit = ONE | TWO | THREE | FOUR | FIVE |
         SIX | SEVEN | EIGHT | NINE | OH | ZERO;
```

Speech Recognition – Parameters (IV)

Ex:

Dial three three two six five four

Dial nine zero four one oh nine

Phone Woodland

Call Steve Young

```
$digit = ONE | TWO | THREE | FOUR | FIVE |
         SIX | SEVEN | EIGHT | NINE | OH | ZERO;
$name  = [ JOOP ] JANSEN |
         [ JULIAN ] ODELL |
         [ DAVE ] OLLASON |
         [ PHIL ] WOODLAND |
         [ STEVE ] YOUNG;
```

Speech Recognition – Parameters (IV)

Ex:

Dial three three two six five four

Dial nine zero four one oh nine

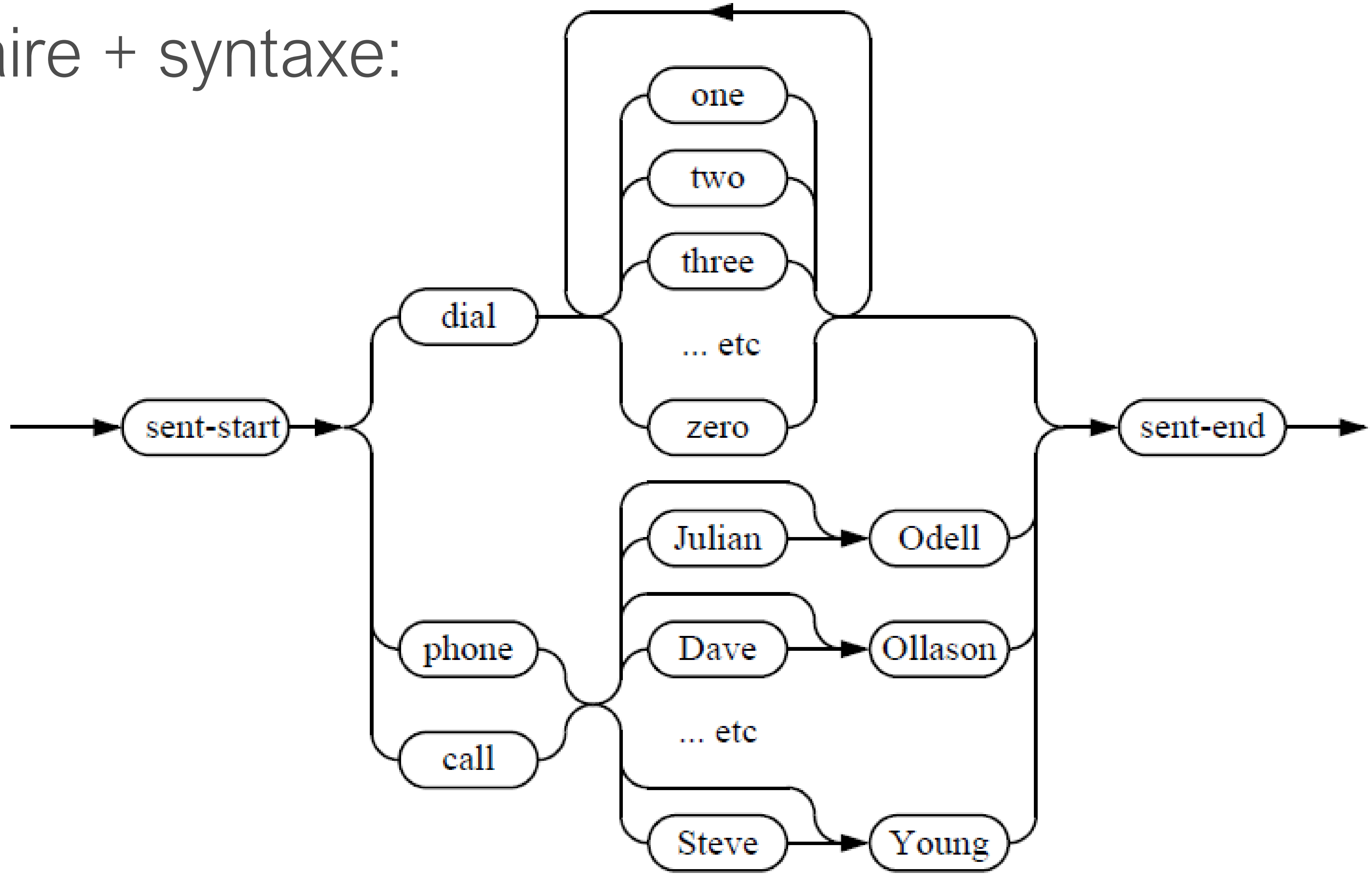
Phone Woodland

Call Steve Young

```
$digit = ONE | TWO | THREE | FOUR | FIVE |
         SIX | SEVEN | EIGHT | NINE | OH | ZERO;
$name  = [ JOOP ] JANSEN |
         [ JULIAN ] ODELL |
         [ DAVE ] OLLASON |
         [ PHIL ] WOODLAND |
         [ STEVE ] YOUNG;
( SENT-START ( DIAL <$digit> | (PHONE|CALL) $name) SENT-END )
```

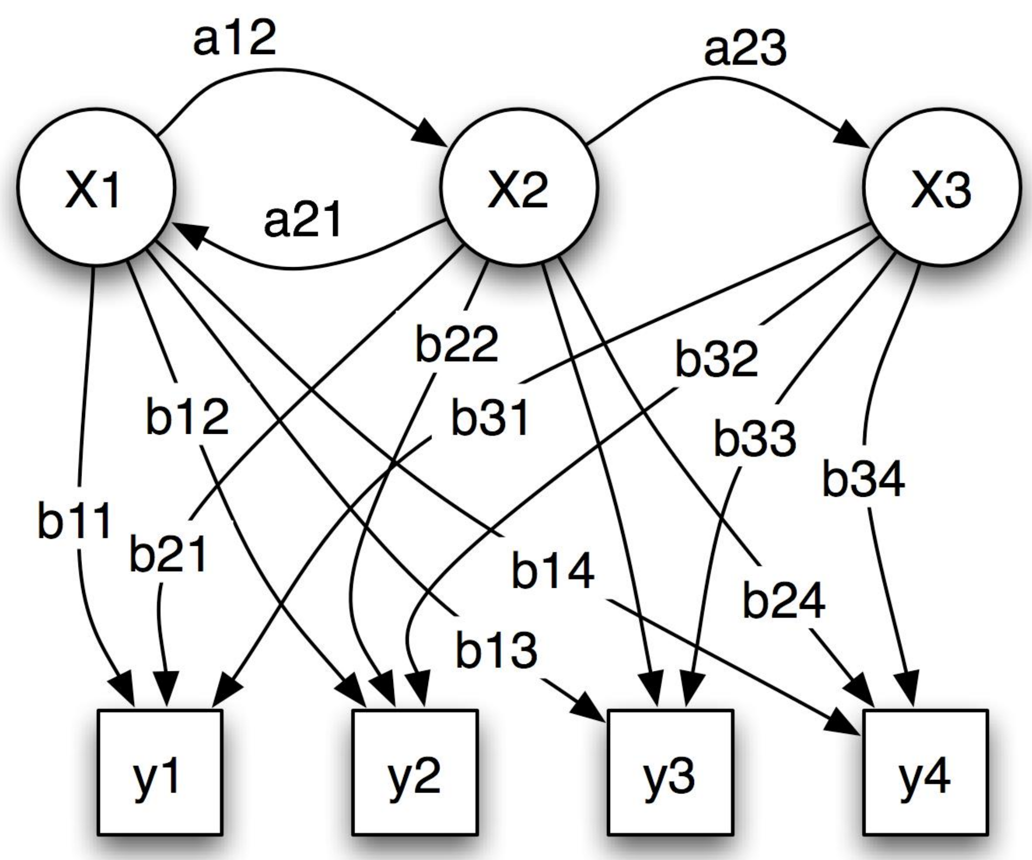
Speech Recognition – Parameters (V)

Grammaire + syntaxe:



Algorithmes – HMMs

- Processing with Hidden Markov Models
  - Advantages
    - User dependent or Independent
    - Not needs to re-train the whole system for small changes
    - Working with time-series
  - Drawbacks
    - Language dependent
    - Can be slow (training)



Part 1 – What you should know

- Classification
  - Definition
- Time series
  - Why are them important?
- Speech processing
  - Motivation
  - Impact of parameter selection

Introduction

HIDDEN MARKOV MODELS

Hidden Markov Models (HMMs)

- Introduction to HMMs
  - Discrete-time Markov chain (*chaînes de Markov à temps discret*)
  - Extension to **H**MMs
  - Emission probabilities (Probabilités d'émission)
  - HMMs elements: $M=(A, B, \pi)$
  - HMMs topologies

HMMs – Introduction

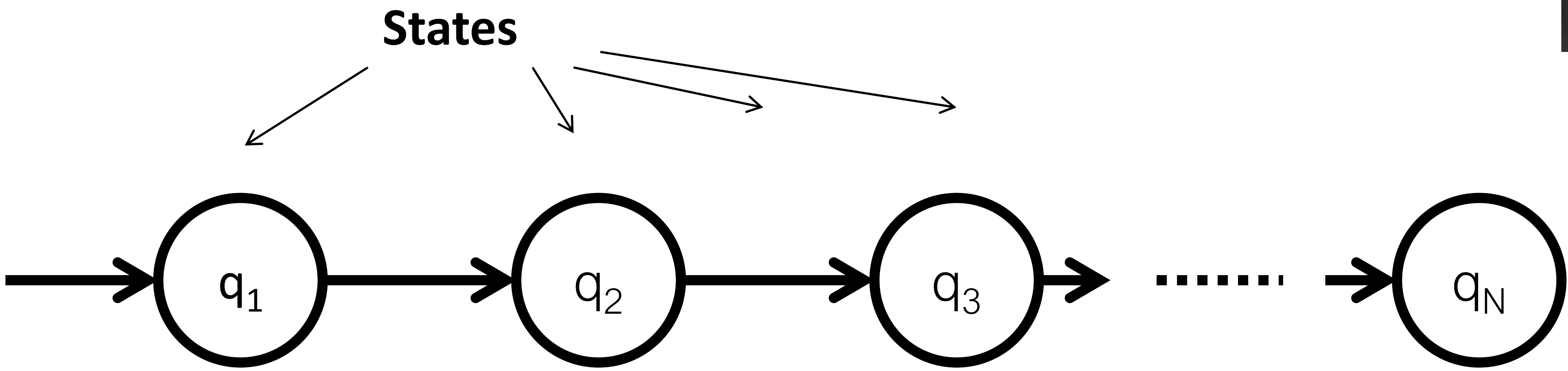
- Theoretical bases published by L.E. Baum in mid ‘60
- First implementation for speech processing in ‘70 @IBM
- Other denominations:
  - « Probabilistic functions of Markov chain »
  - « Markov sources »
- The HMMs can be seen as extensions of Markov chains

HMMs – Introduction

- Nowadays used in very different domains:
  - Automatic speech recognition
  - Handwriting recognition
  - Biometrics: speaker and writer verification
  - Bioinformatics: search in DNA sequences, protein modeling
  - Linguistic: word modeling
  - Anomaly detection
  - …

(H)MMs – Introduction

- Markov model:
  - *“The future is independent from the past given the present”*



Discrete-time Markov chain (I)

- Consider a system that is in a state *i* from a set of *N* states
- The system changes state at each discrete time according to a set of transition probabilities associated with each state
- Markov stochastic process of order 1: process without memory, i.e.:

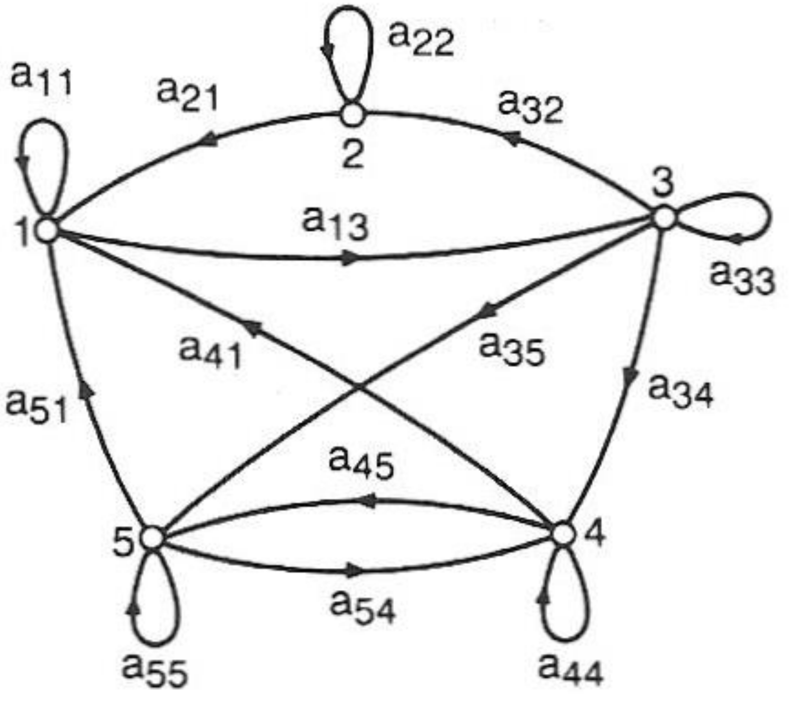


Figure 6.1 A Markov chain with five states (labeled 1 to 5) with selected state transitions.

$$P(q_t = j \mid q_{t-1} = i, q_{t-2} = k, \dots) = P(q_t = j \mid q_{t-1} = i) = a_{ij}$$

with

$$\sum_{j=1}^{N} a_{ij} = 1$$

Discrete-time Markov chain (II)

- Example: Markov model of measured weather data
  - State 1 = rainy
  - State 2 = cloudy
  - State 3 = sunny

- What is the probability of a sequence of days: sunny-sunny-sunny-rain-rain-cloudy-sunny?

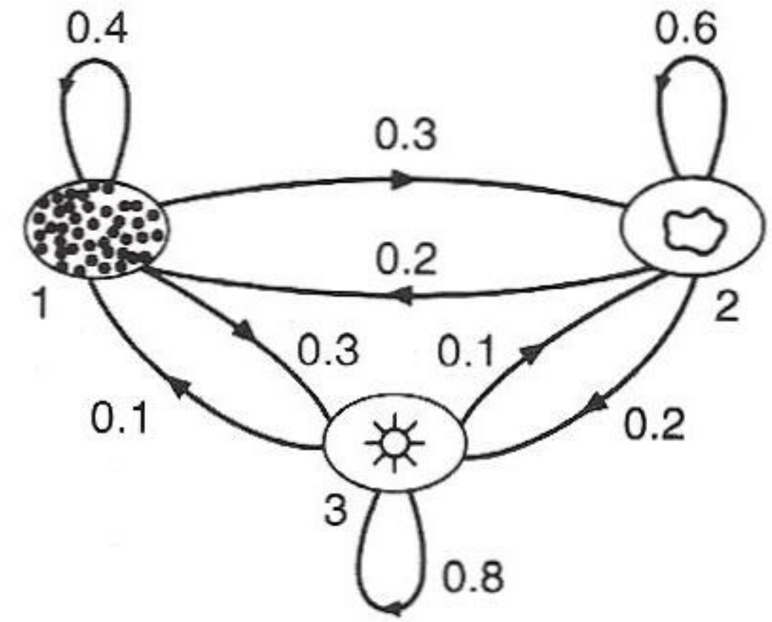


Figure 6.2 Markov model of the weather.

Discrete-time Markov chain (II)

- Example: Markov model of measured weather data
  - State 1 = rainy
  - State 2 = cloudy
  - State 3 = sunny

- What is the probability of a sequence of days: sunny-sunny-sunny-rain-rain-cloudy-sunny?

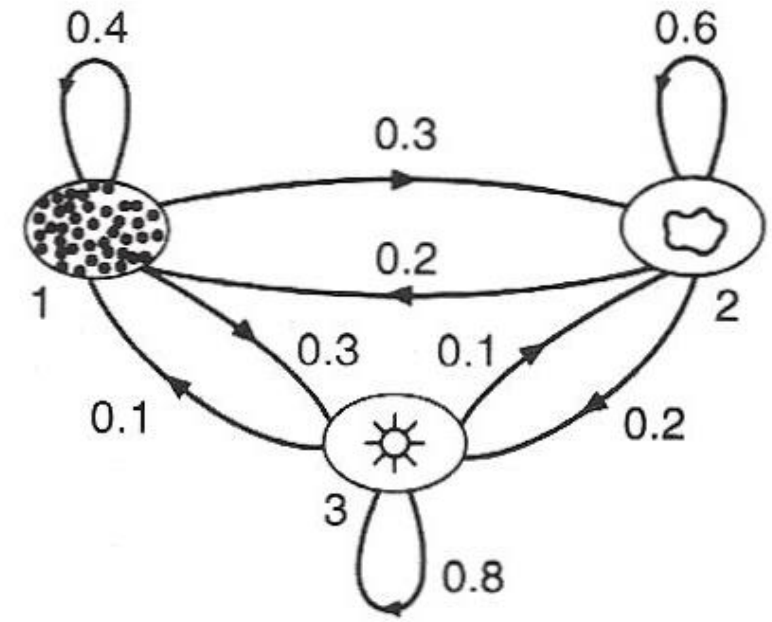


Figure 6.2 Markov model of the weather.

$$P(X \mid Model) = P(s, s, s, \mathrm{r}, \mathrm{r}, \mathrm{c}, s \mid Model)$$

Discrete-time Markov chain (II)

- Example: Markov model of measured weather data
  - State 1 = rainy
  - State 2 = cloudy
  - State 3 = sunny

- What is the probability of a sequence of days: sunny-sunny-sunny-rain-rain-cloudy-sunny?

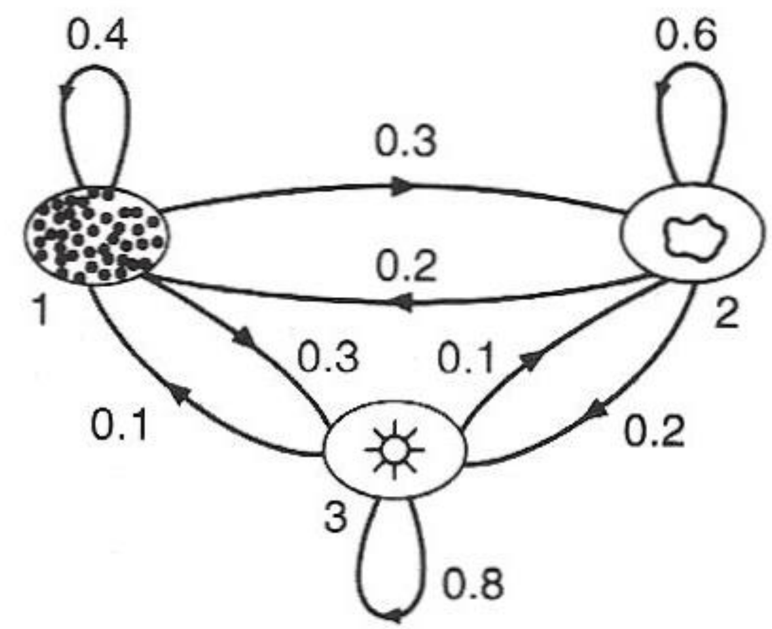


Figure 6.2 Markov model of the weather.

$$
\begin{aligned}
P(X \mid Model) &= P(s, s, s, \mathrm{r}, \mathrm{r}, \mathrm{c}, s \mid Model) \\
&= P(s \mid s)P(s \mid s)P(\mathrm{r} \mid s)P(\mathrm{r} \mid \mathrm{r})P(\mathrm{c} \mid \mathrm{r})P(s \mid \mathrm{c}) \\
&= a_{33}a_{33}a_{31}a_{11}a_{12}a_{23} \\
&= (0.8)^2(0.1)(0.4)(0.3)(0.2) \\
&= 1.536 \times 10^{-3}
\end{aligned}
$$

Discrete-time Markov chain (II)

- Example: Markov model of measured weather data
  - State 1 = rainy
  - State 2 = cloudy
  - State 3 = sunny

- What is the probability of a sequence of days: sunny-sunny-sunny-rain-rain-cloudy-sunny?

Note: in addition to the transition probabilities, the model can also define the **probability of the initial state π_i** (it defines the probability of starting in state i)

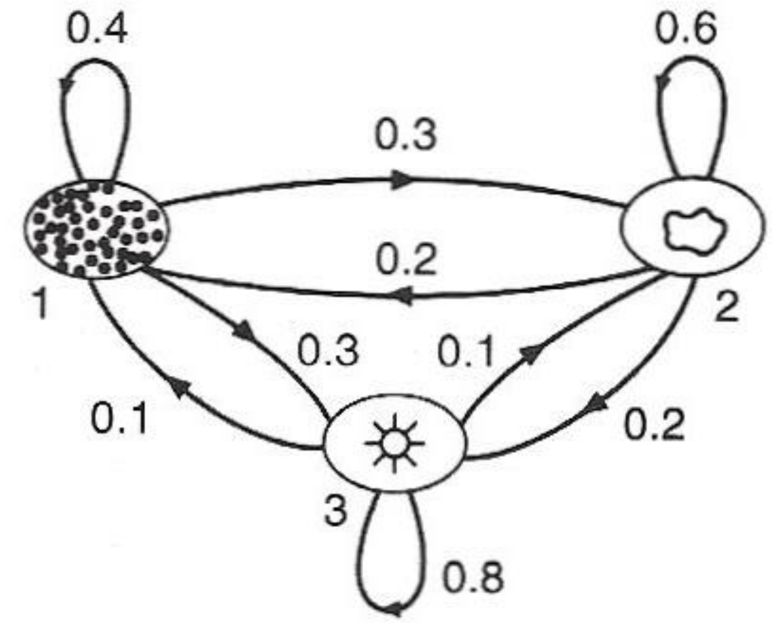
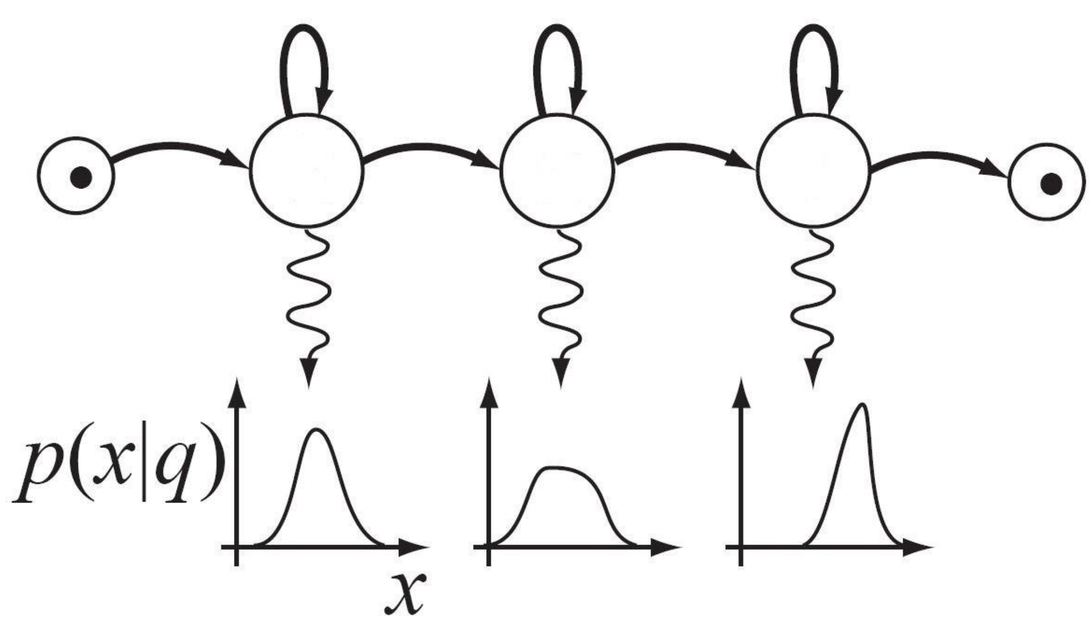


Figure 6.2 Markov model of the weather.

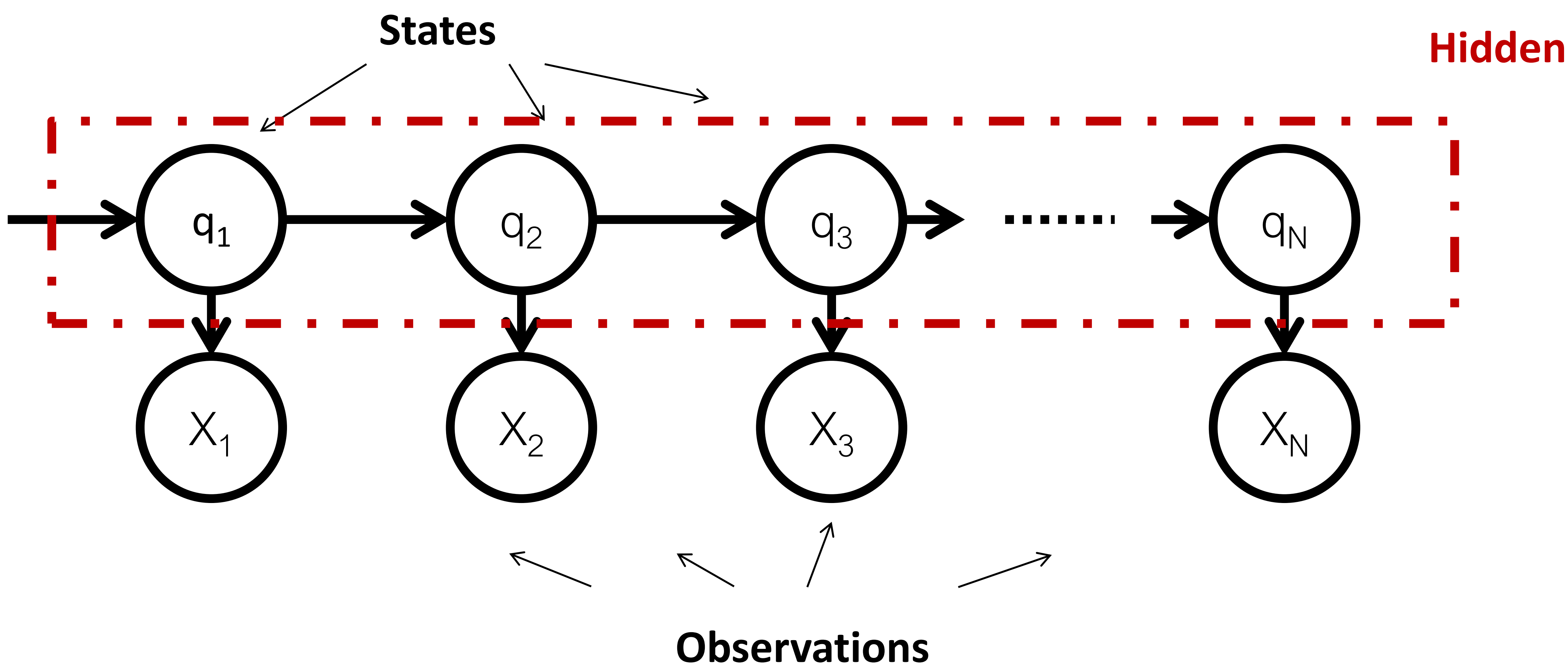
$$
\begin{aligned}
P(X \mid Model) &= P(s, s, s, \mathrm{r}, \mathrm{r}, \mathrm{c}, s \mid Model) \\
&= P(s \mid s) P(s \mid s) P(\mathrm{r} \mid s) P(\mathrm{r} \mid \mathrm{r}) P(\mathrm{c} \mid \mathrm{r}) P(s \mid \mathrm{c}) \\
&= a_{33} a_{33} a_{31} a_{11} a_{12} a_{23} \\
&= (0.8)^2 (0.1)(0.4)(0.3)(0.2) \\
&= 1.536 \times 10^{-3}
\end{aligned}
$$

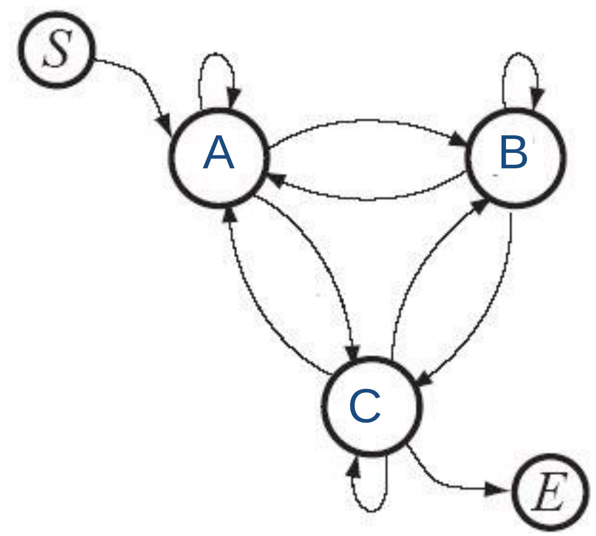
Extension to **Hidden** Markov models

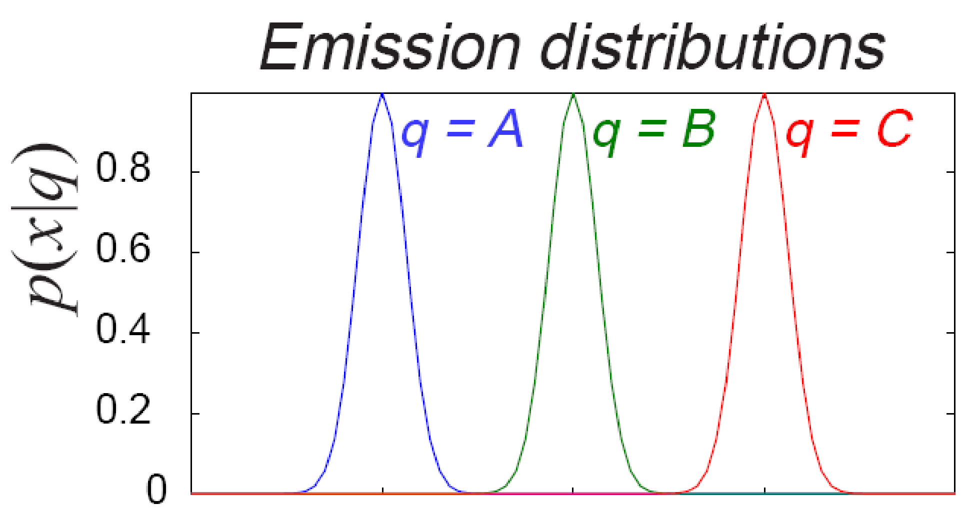
- An observation becomes a probabilistic function dependent on the state
- In other words, the observation x has a certain probability of being “emitted" in a state q
- It is called *emission probability* $p(x|q)$

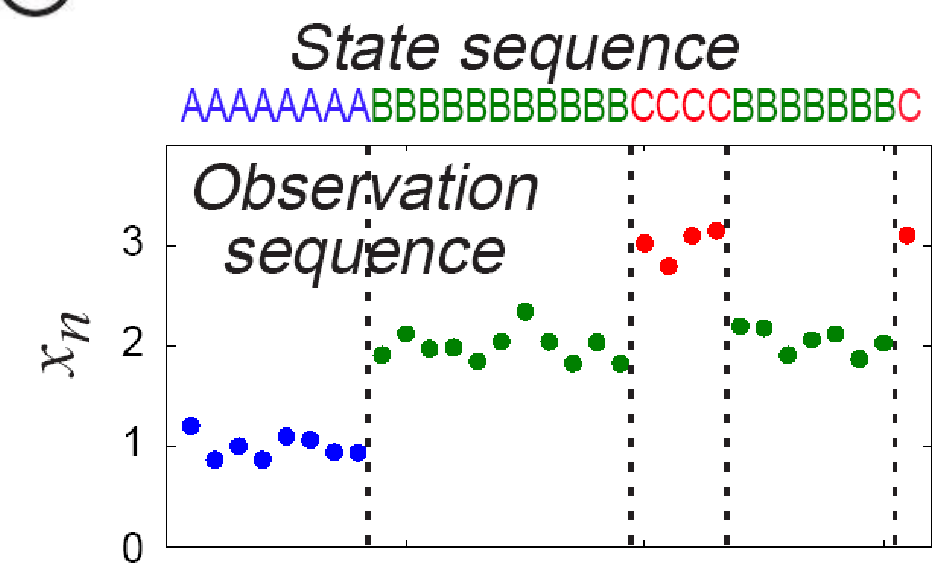


Extension to Hidden Markov Models

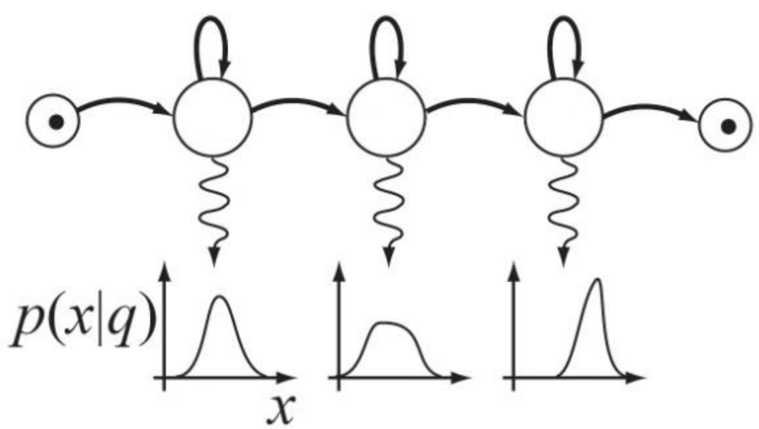


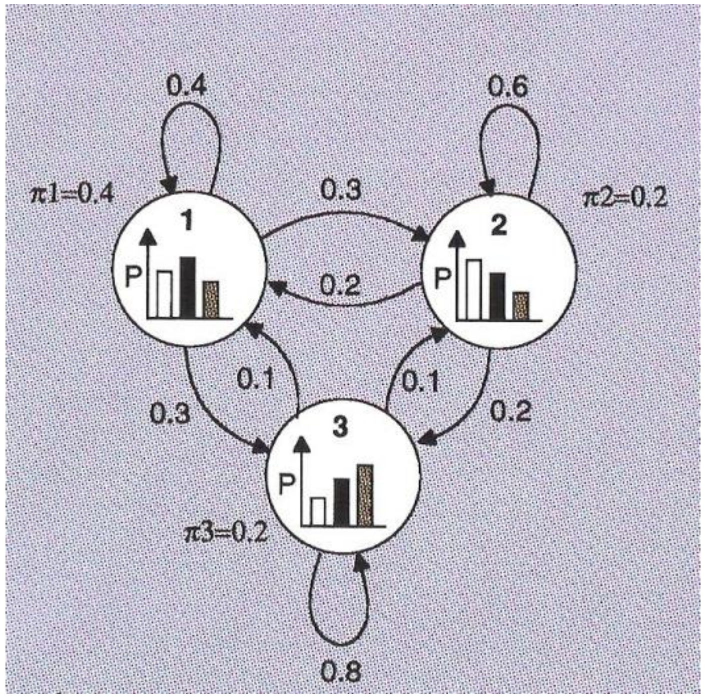

S
A
B
C
E


Emission distributions
p(x|q)
0.8
0.6
0.4
0.2
0
q = A
q = B
q = C


State sequence
AAAAAAAABBBBBBBBBBBCCCCBBBBBBBC
Observation sequence
x_n
3
2
1
0

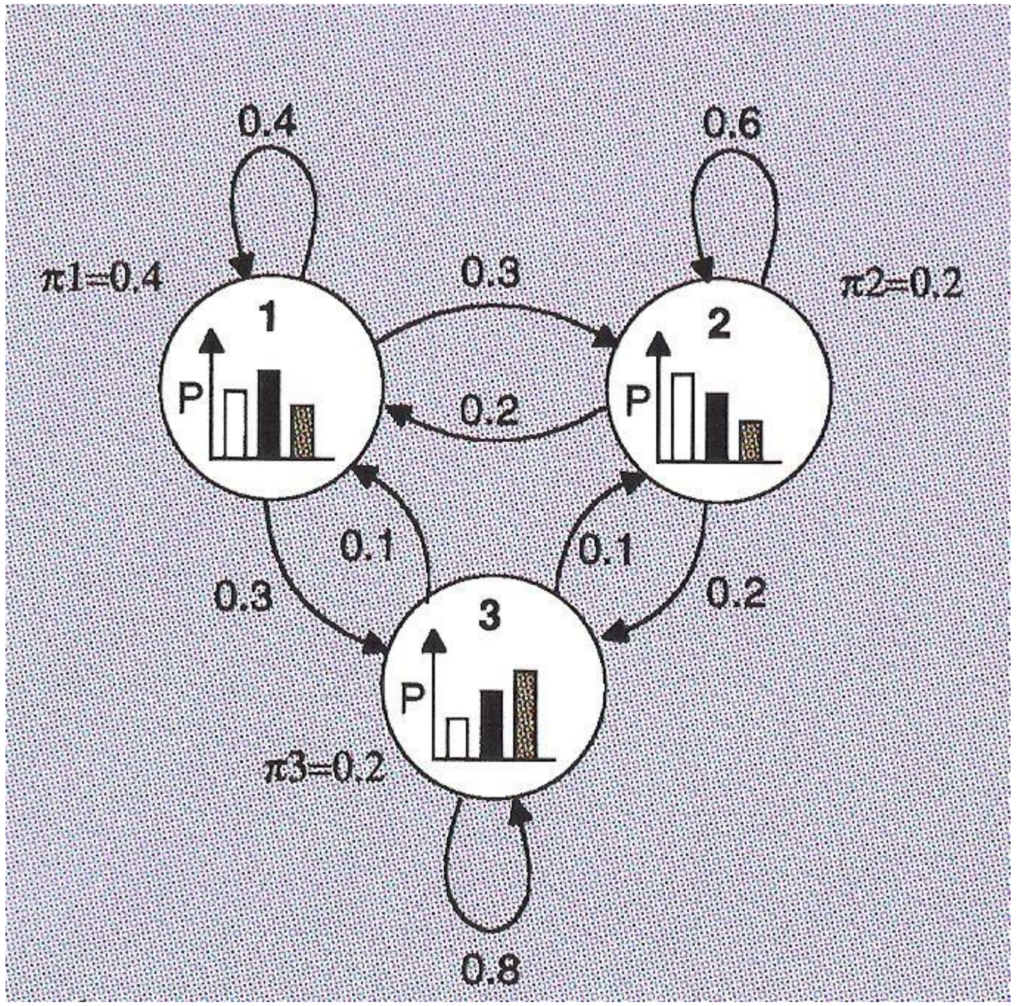
Emission probabilities





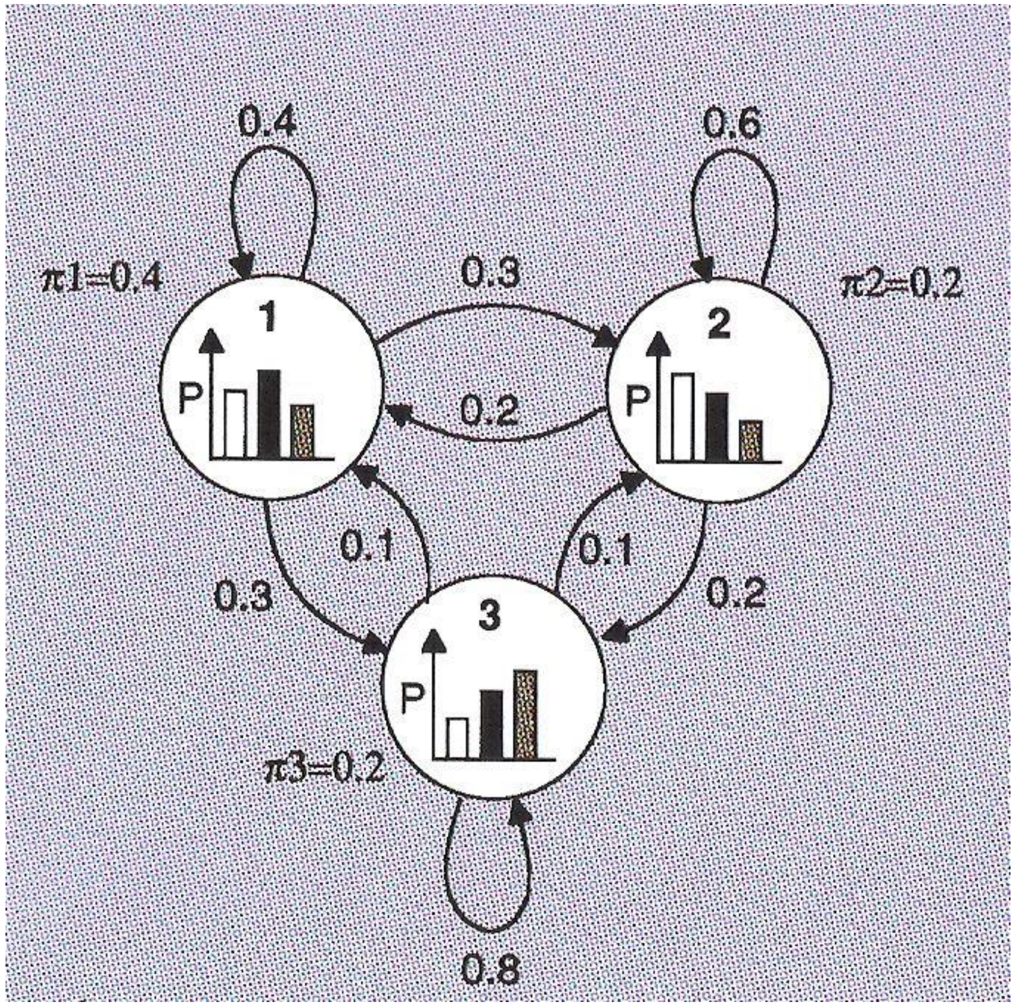
- Observations are continuous:
  - $p(x|q)$ follows a continuous distribution: Gaussian, multi-Gaussian, …
- Observations are discrete:
  - $p(x|q)$ is modeled by an histogram

Emission probabilities



- *A*, the set of transition probabilities
- *B*, the set of emission probabilities density
- π, the set of initial state probabilities
- $M=(A, B, \pi)$

Emission probabilities



$$q_t = \text{state at time } t$$

$$A = \{a_{ij}\} \quad \text{with} \quad a_{ij} = P(q_t = j \mid q_{t-1} = i)$$

$$B = \{b_j(x)\} \quad \text{with} \quad b_j(x) = P(x \mid q_t = j)$$

$$\pi = \{\pi_i\} \quad \pi_i = P(q_1 = i)$$

$$M = (A, B, \pi)$$

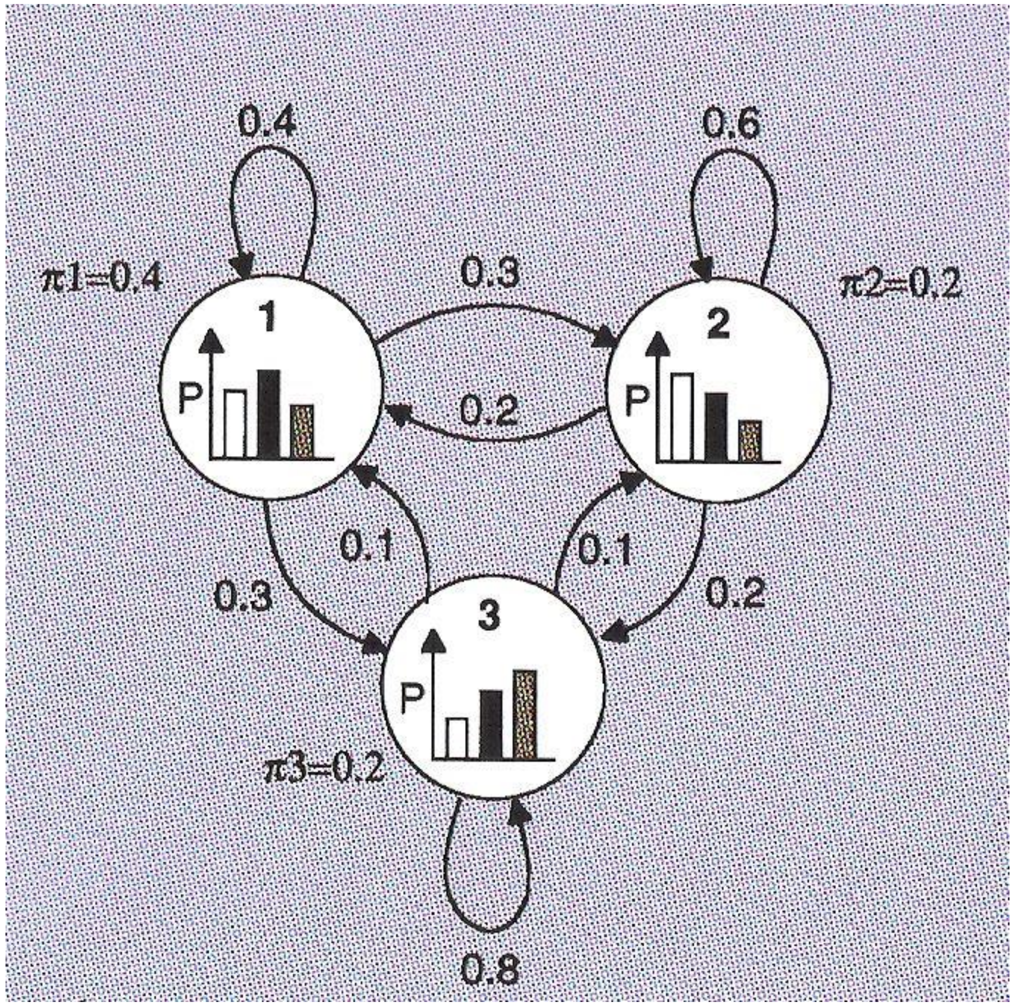
Sequence of observations

$$X = \{x_1, \ldots, x_n\}$$

Particular sequence of states

$$\mathrm{q} = \{q^1, \ldots, q^n\}$$

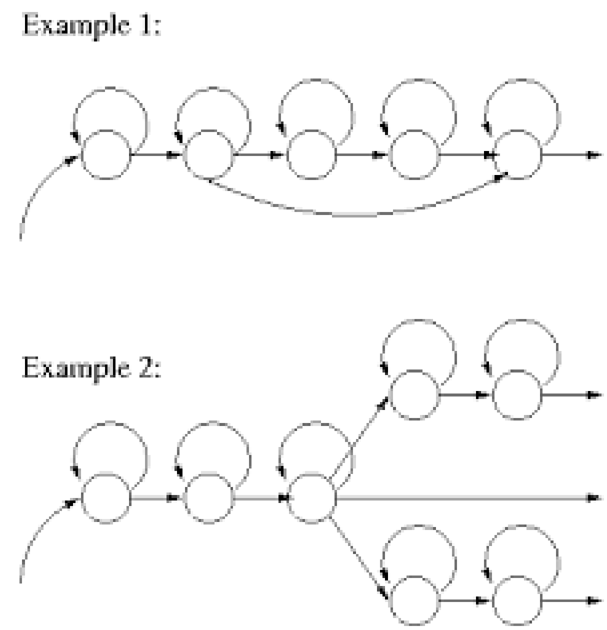
Quiz: Find the error

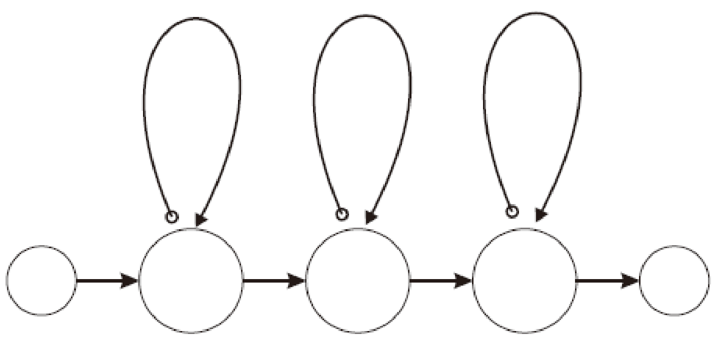


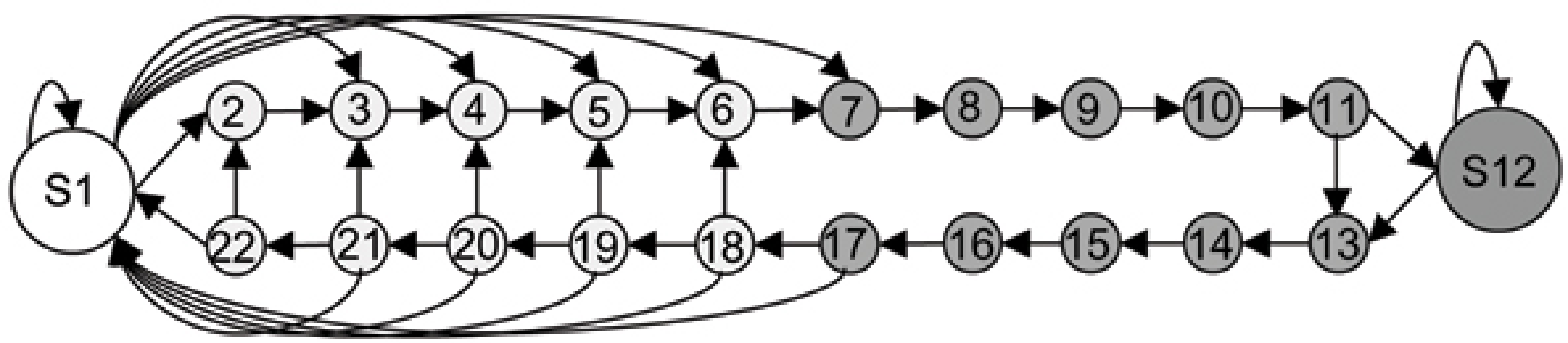
HMMs topologies

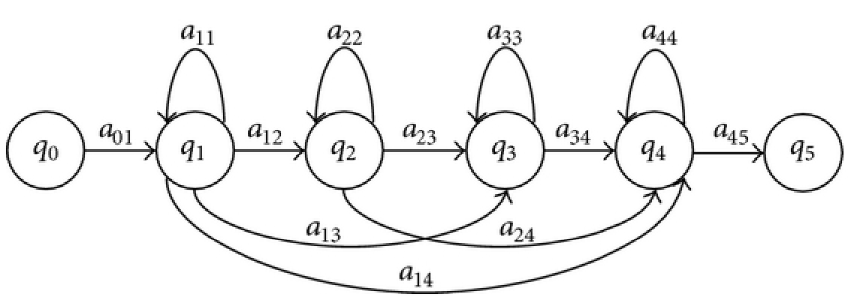
- The structure (or topology) of a HMM can be learnt in the training phase
- However, imposing a “correct” topology from the beginning can hugely speed-up the training phase (less data required)

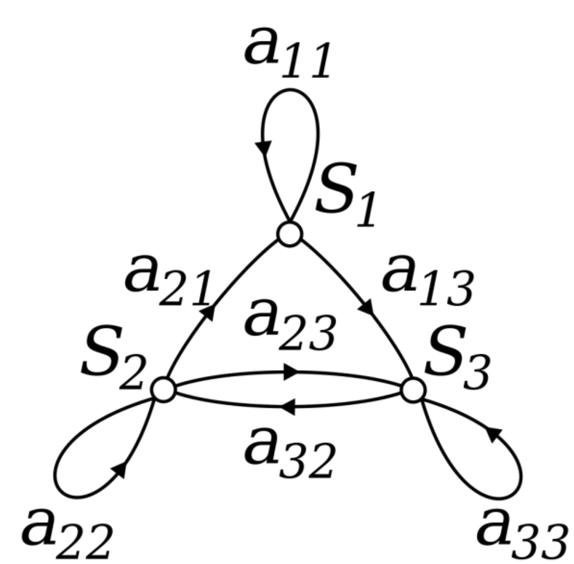
How to chose?


Example 1:
Example 2:

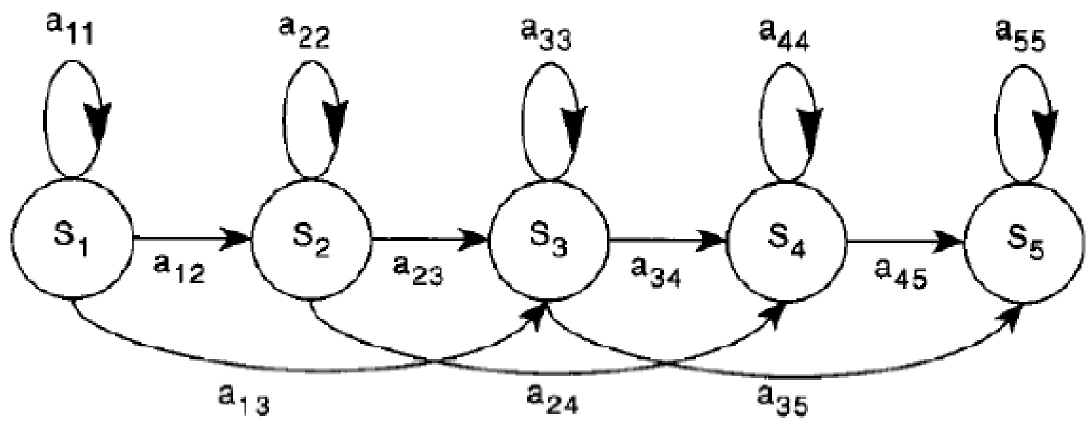



S1
2
3
4
5
6
7
8
9
10
11
S12
13
14
15
16
17
18
19
20
21
22

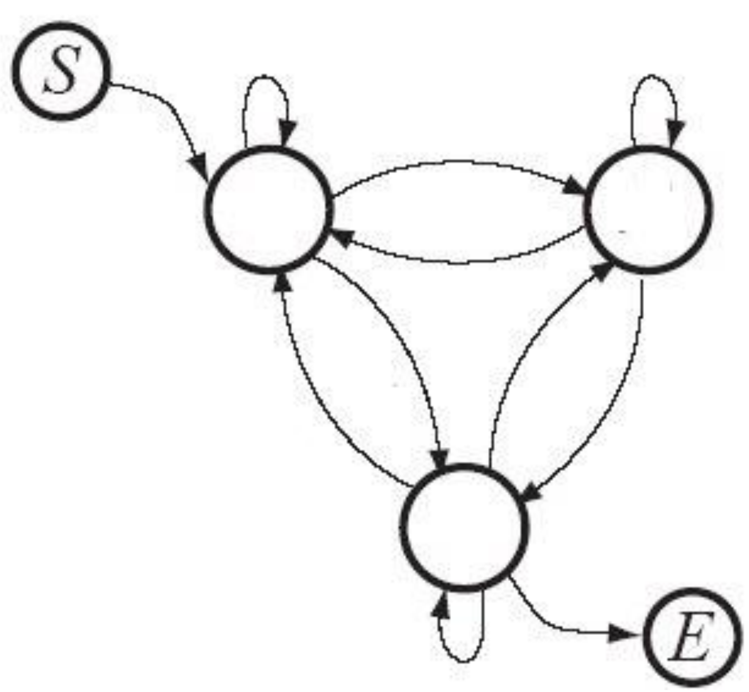

a_{11}
a_{22}
a_{33}
a_{44}
q_0
a_{01}
q_1
a_{12}
q_2
a_{23}
q_3
a_{34}
q_4
a_{45}
q_5
a_{13}
a_{24}
a_{14}


a_{11}
S_1
a_{21}
a_{13}
a_{23}
S_2
S_3
a_{32}
a_{22}
a_{33}

HMMs main topologies



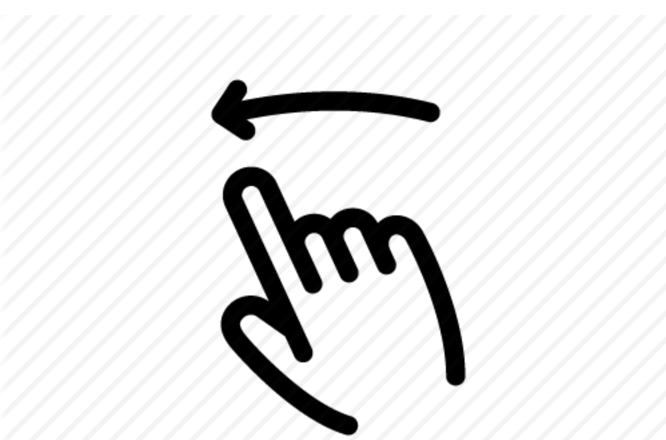
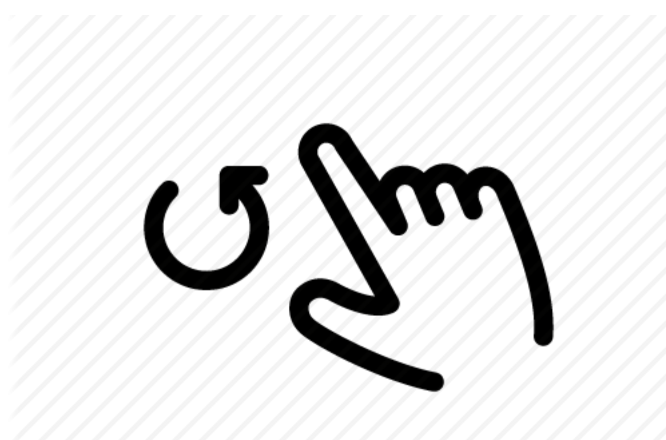
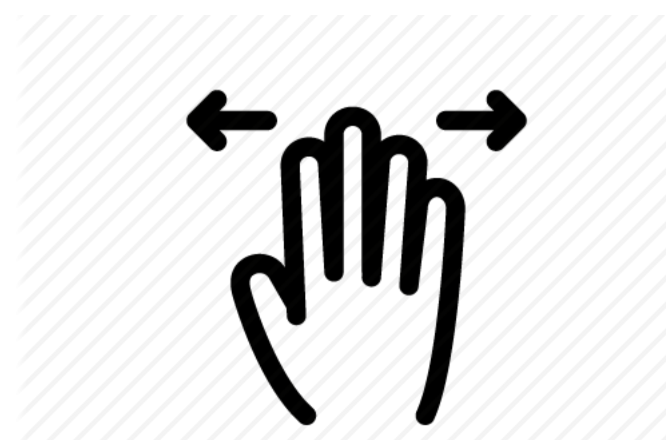
Left-Right (LR)



Ergodic

Exercise: HMMs topologies

- Define the model for 3 gestures
  - Swipe
  - Circle
  - Waving
- Imagine you have accelerometer data acquired at 100Hz

How to use HMM in practice

What you will do in the next practical session

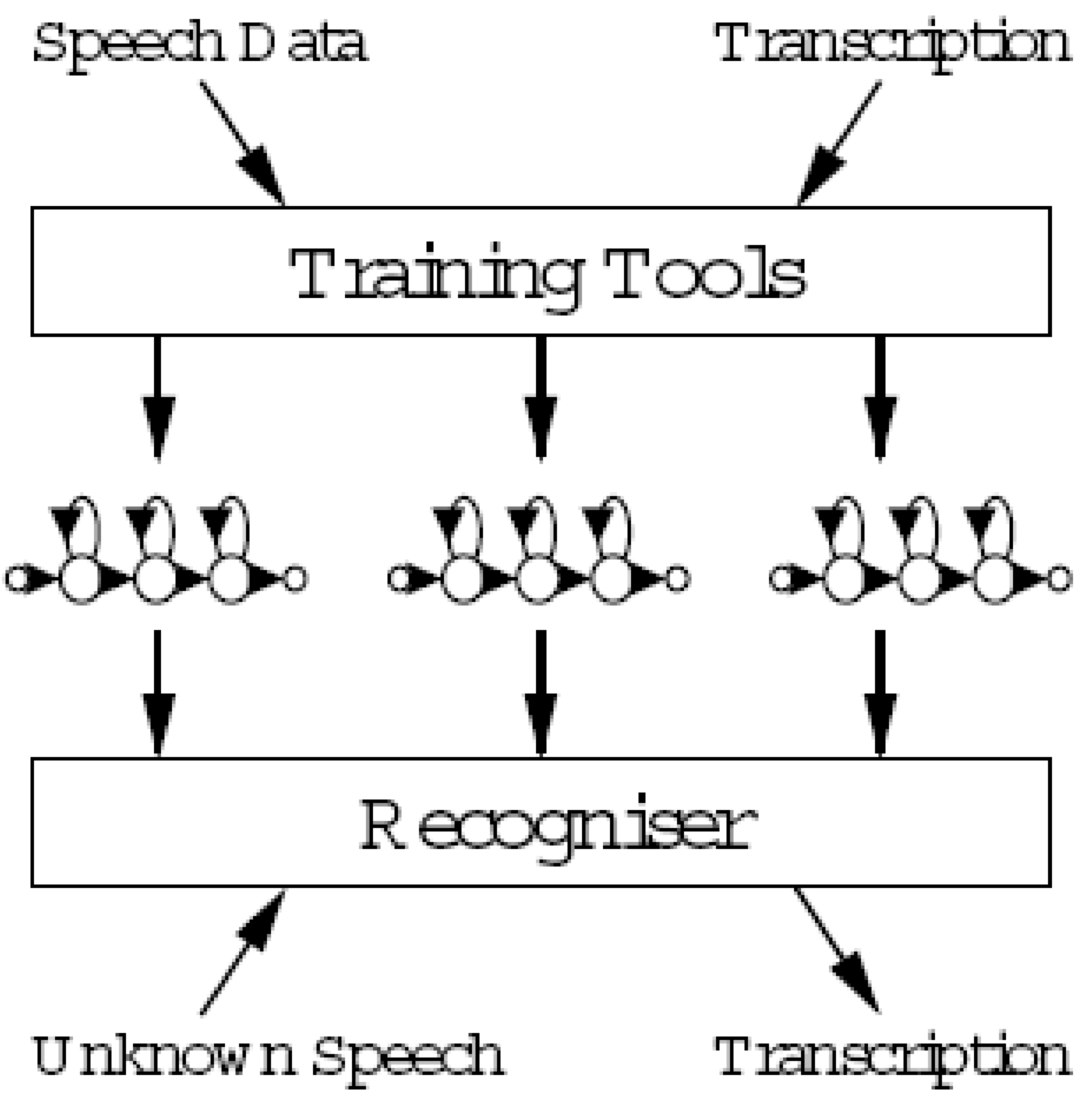
APPLICATIONS

Classification Steps using HMMs

1. Acquire data (!)
2. Create a model for your problem
   - In particular, 1 model for each class
   - Define the topology, the initial parameters, etc.
   - Simplification: similar models for each class (different number of states but same topology: ergodic, left-right)
3. Train your model
   - Compute values of A, B, π, M
4. Validate your model
5. Improve the model (iterate 3. and 4.)
6. Test your model and asses the quality of your models

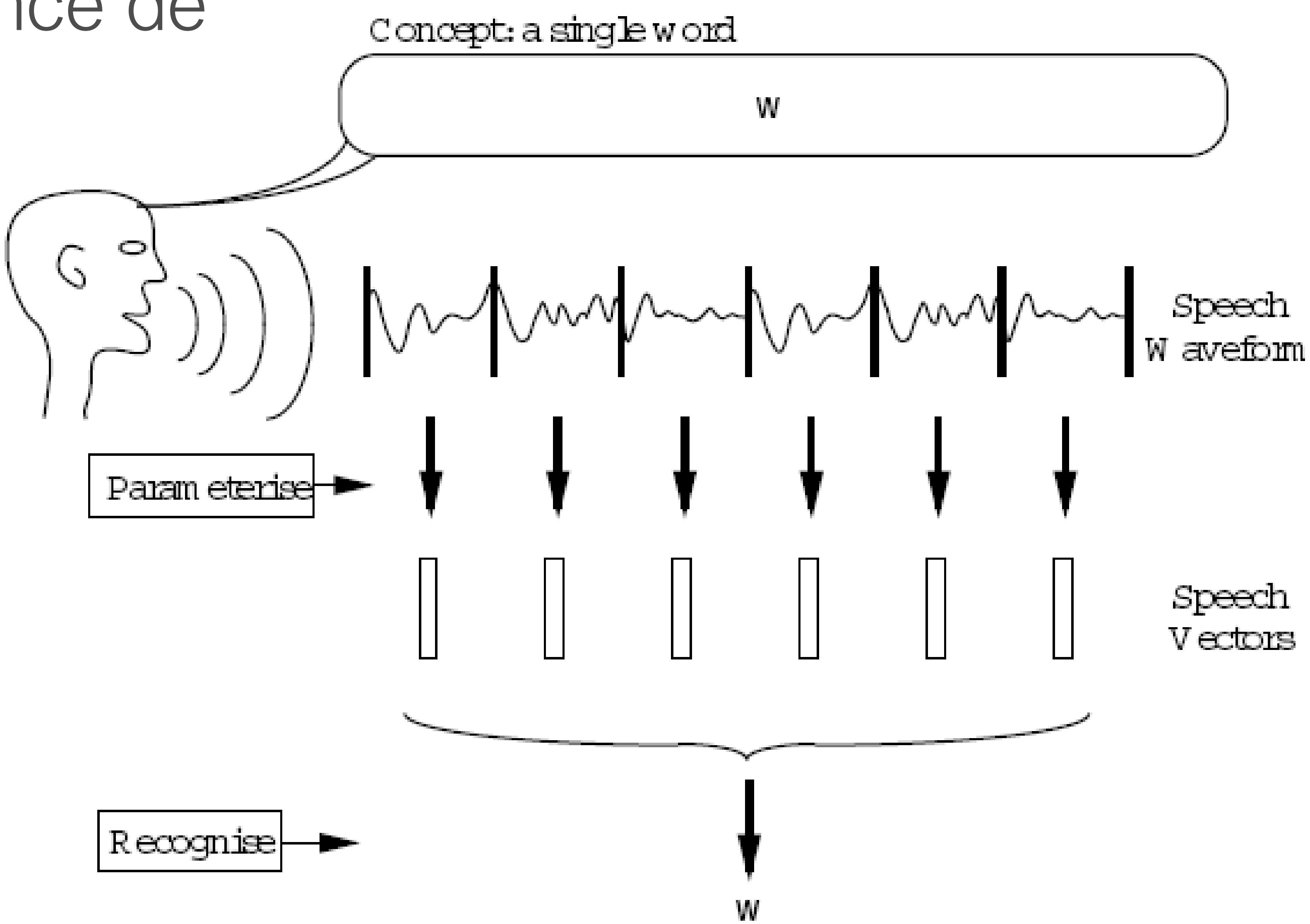
Reconnaissance de la parole

- Vue d'ensemble



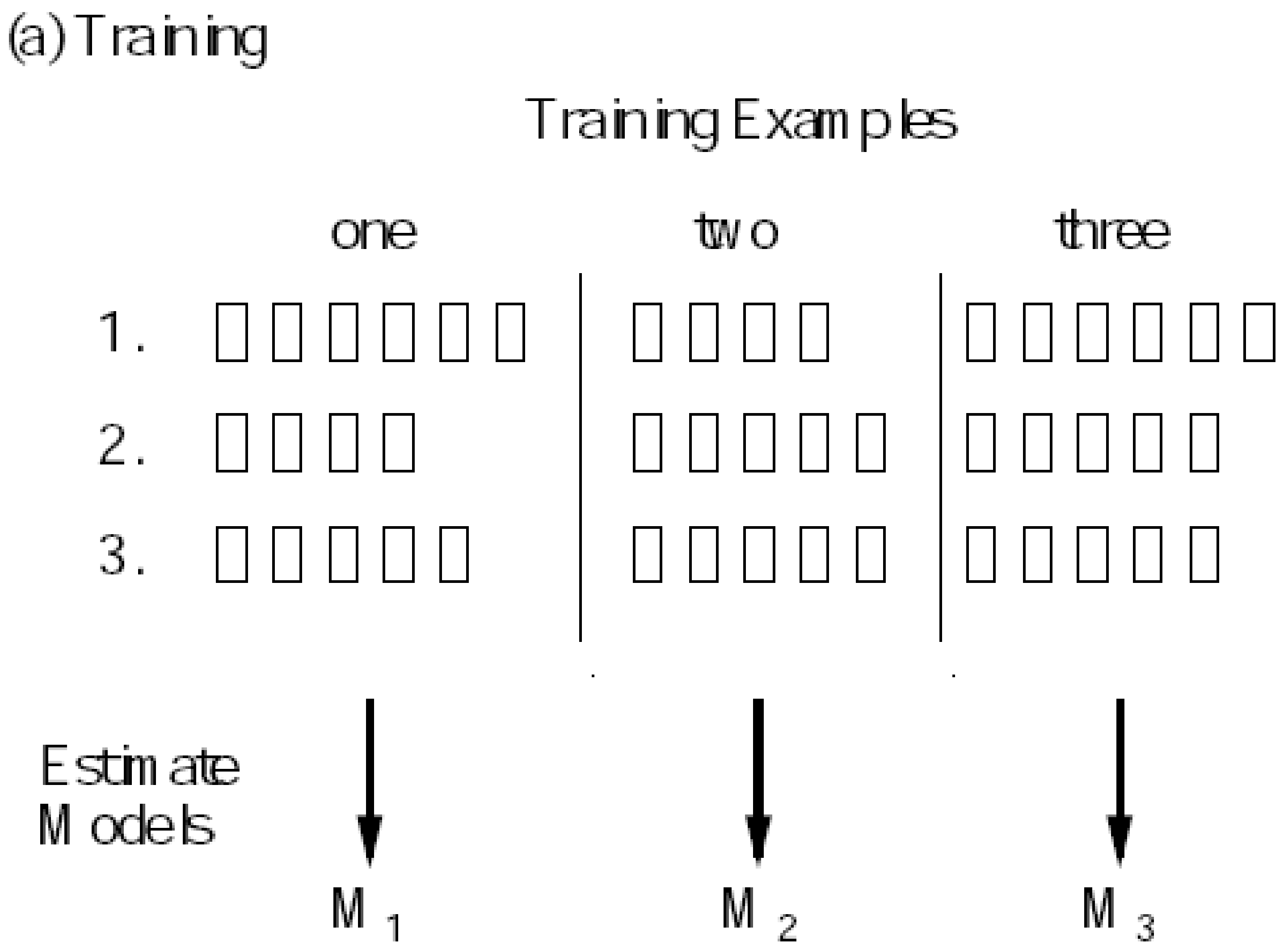
Reconnaissance de la parole

- Reconnaissance de mots isolés



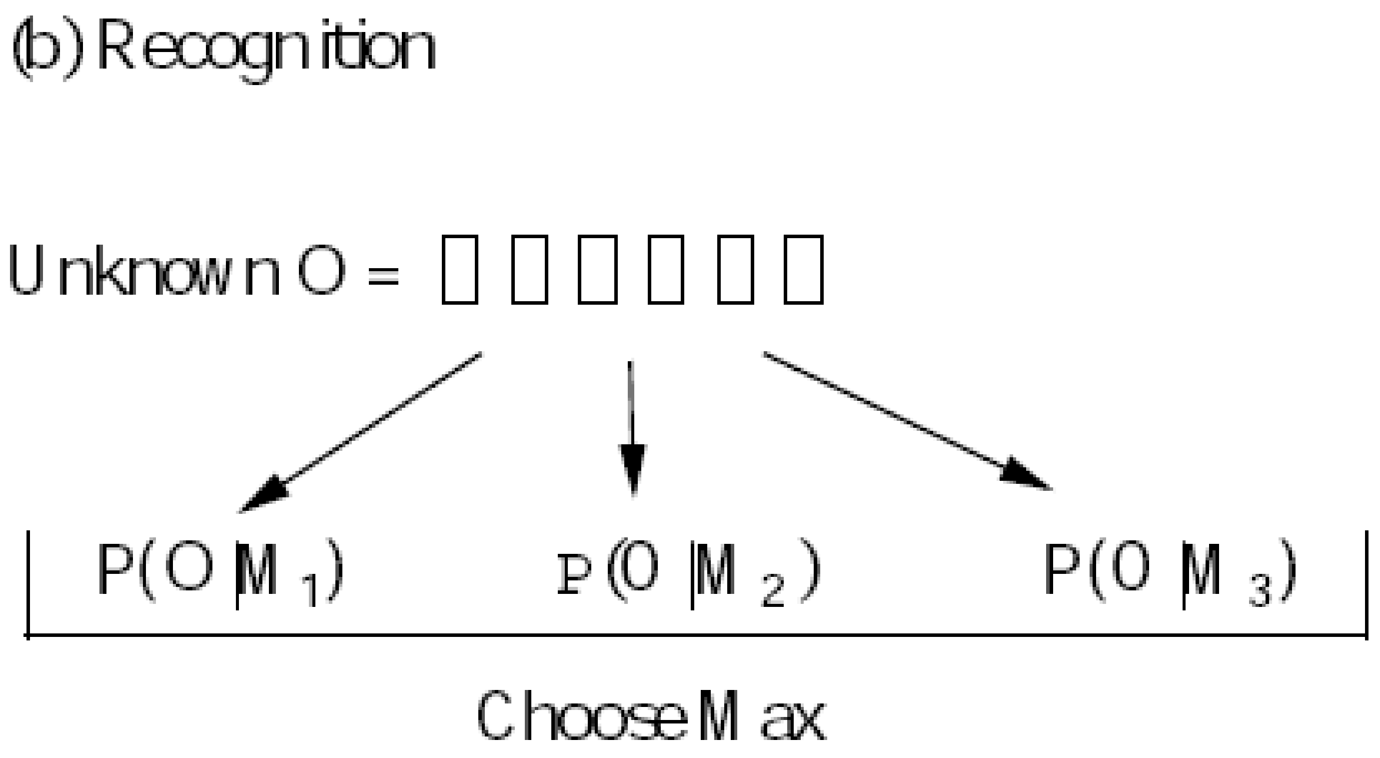
Reconnaissance de la parole

- Training
  - Estimation des modèles



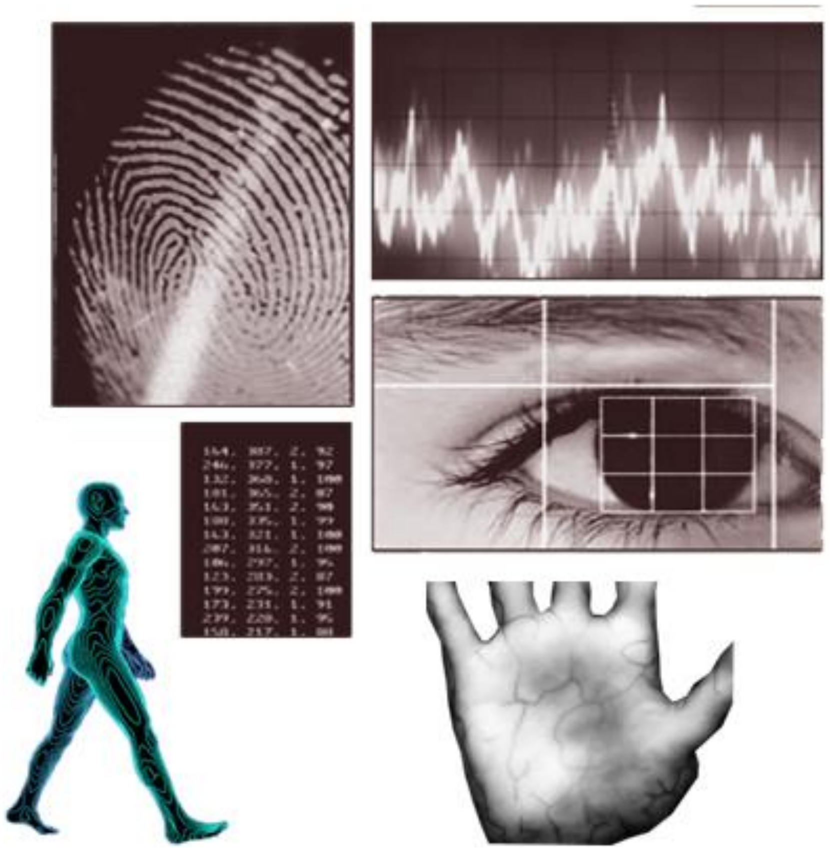
Reconnaissance de la parole

- Reconnaissance



Other application: biometry

- Binary problem
  - (Yes => access granted; No => access denied)



Other application: biometry

- Speaker identification

Homeworks!

Before the next session:

- Read the TP (Moodle)
- Record the dataset

Part 2 - What you should know

- Hidden Markov Models
  - Markov Models Vs Hidden Markov Models
- HMM parameters
  - A, B, π, M
- Why HMMs are relevant dealing with time series?
- The application of HMM in speech processing (see the TP)

References

[1] L. R. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition,” *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[2] HTKBook - http://htk.eng.cam.ac.uk/docs/docs.shtml

ML for TimeSeries alternative to HMM

- Neural Networks!
  - Recurrent Neural Networks
    - LSTM – GRU
- Convolutional Neural Networks
  - 1D convolutions
  - TCN (Temporal Convolutional Networks)