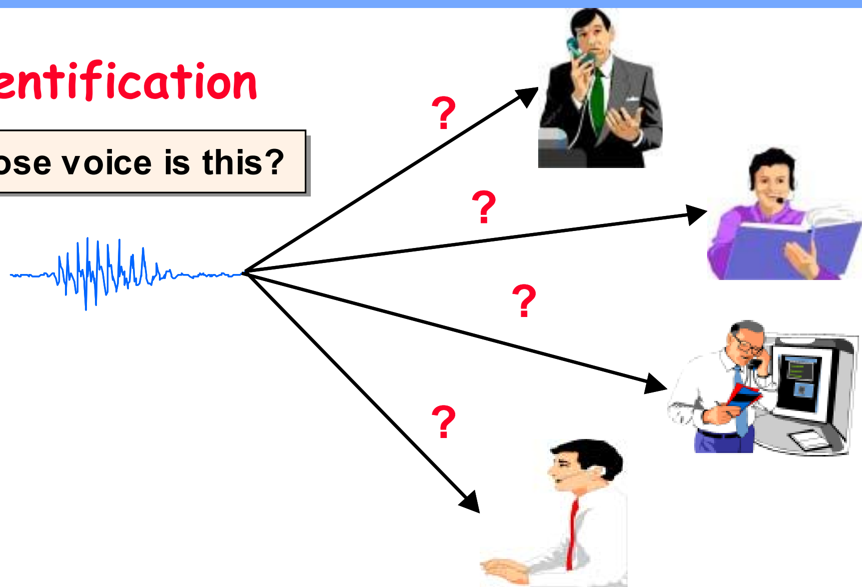
Three Speaker Recognition Tasks

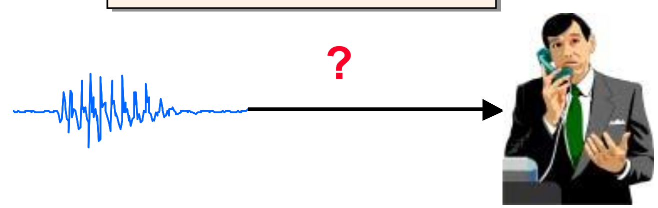
Identification

Whose voice is this?



Verification/Authentication/Detection

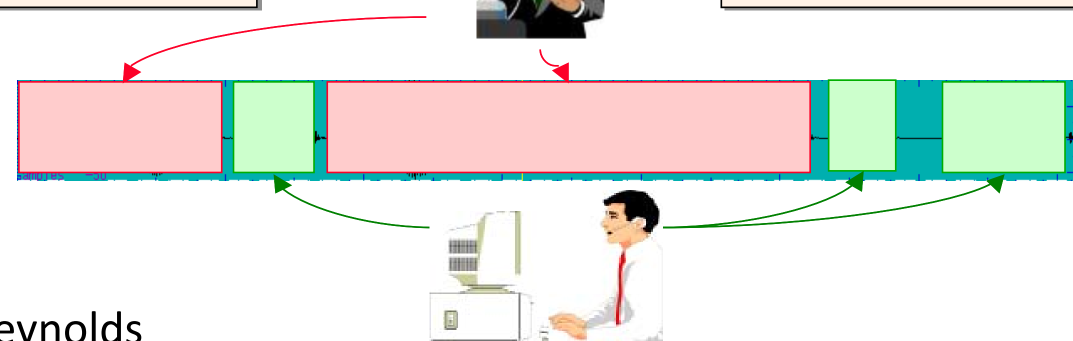
Is this Bob's voice?



Segmentation and Clustering (Diarization)

Where are speaker changes?

Which segments are from the same speaker?

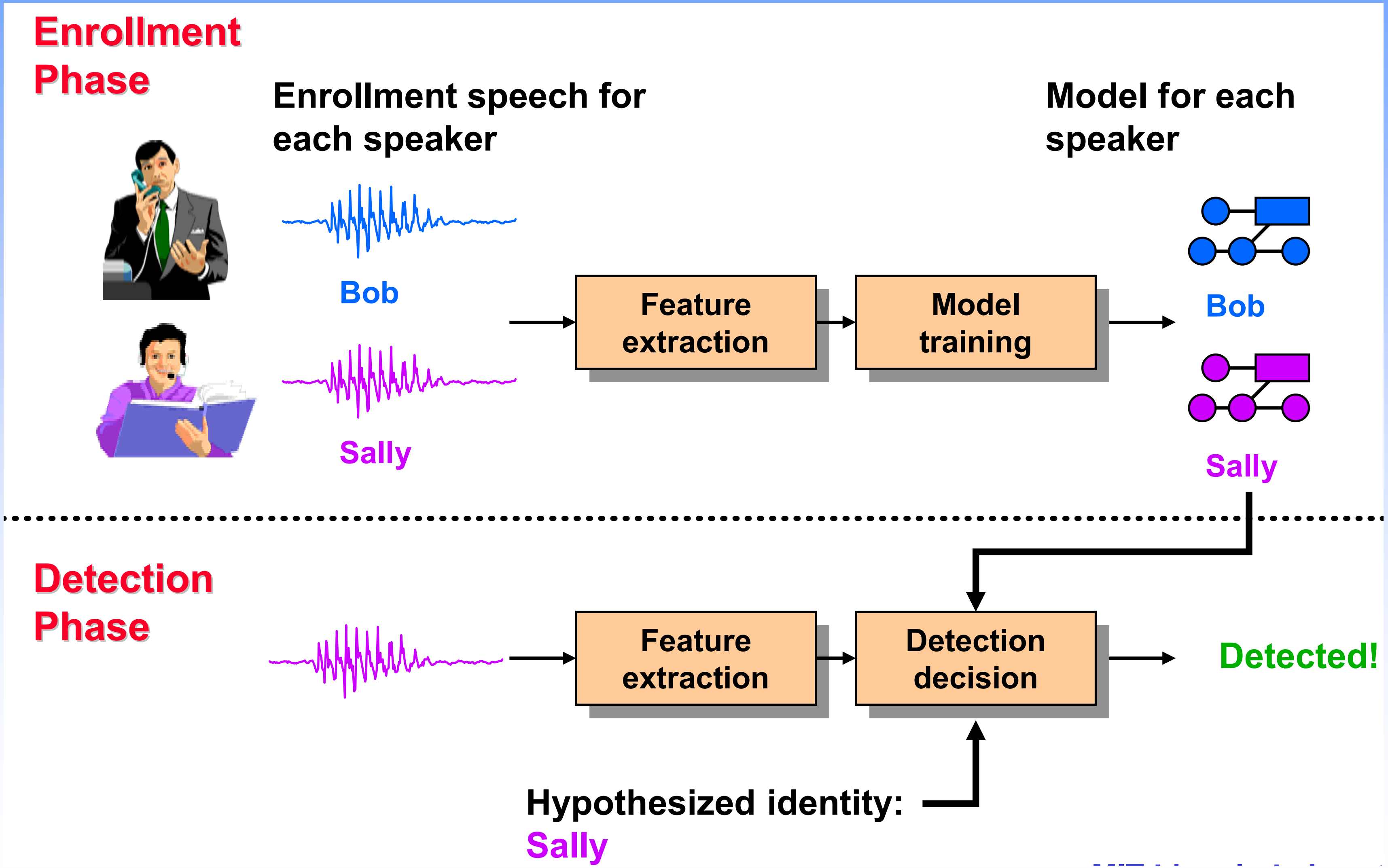


slide from Douglas Reynolds

Two kinds of speaker verification

- Text-dependent
  - Users have to say something specific
  - easier for system
- Text-independent
  - Users can say whatever they want
  - more flexible but harder

Two phases to speaker detection



slide from Douglas Reynolds

Detection: Likelihood Ratio

- Two-class hypothesis test:

  H0: X is **not** from the hypothesized speaker

  H1: X is from the hypothesized speaker
- **Choose the most likely hypothesis**

$$\Pr(H1 \mid X) \underset{<}{\overset{>}{}} \Pr(H0 \mid X)$$

$$\frac{p(X \mid H1)\Pr(H1)}{p(X)} \underset{<}{\overset{>}{}} \frac{p(X \mid H0)\Pr(H0)}{p(X)}$$

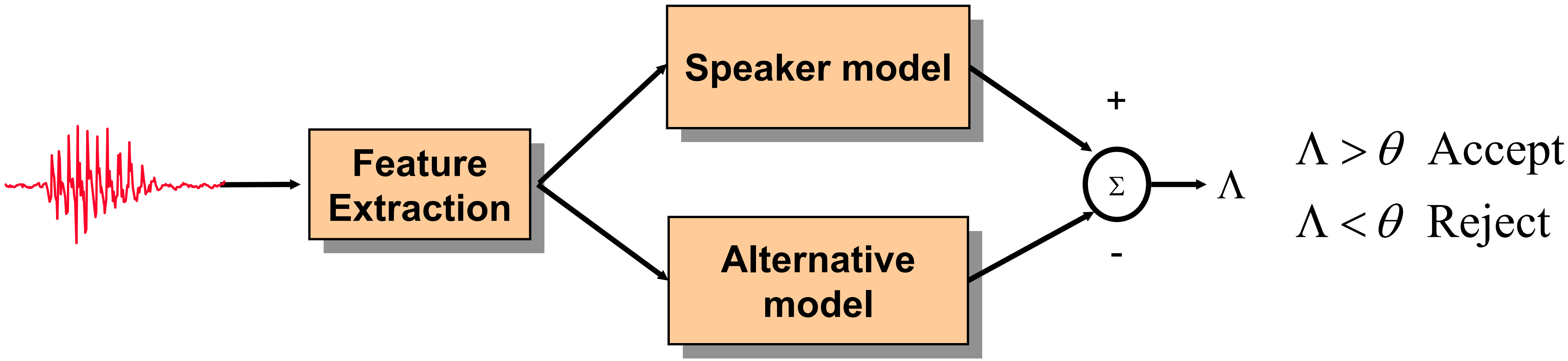
$$\frac{p(X \mid H1)}{p(X \mid H0)} \underset{<}{\overset{>}{}} \frac{\Pr(H0)}{\Pr(H1)}$$

- **Likelihood ratio test:**

$$LR = \frac{p(X \mid H1)}{p(X \mid H0)} \qquad \begin{array}{ll} LR > \theta & \text{Accept } H1 \\ LR < \theta & \text{Accept } H0 \end{array}$$

Speaker ID
Log-Likelihood Ratio Score

LLR= Λ =log *p*(X|H1) – log *p*(X|H0)



- Need *two* models
  - Hypothesized speaker model for H1
  - Alternative (background) model for H0

slide from Douglas Reynolds

How do we get H1?

- Pool speech from several speakers and train a single model:
  - a universal background model (UBM)
  - can train one UBM and use as H1 for all speakers
  - Should be trained using speech representing the expected impostor speech
  - Same type speech as speaker enrollment (modality, language, channel)

Slide adapted from Chu, Bimbot, Bonastre, Fredouille, Gravier, Magrin-Chagnolleau, Meignier, Merlin, Ortega-Garcia, Petrovska-Delacretaz, Reynolds

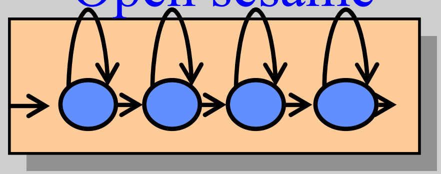
How to compute P(H|X)?

- Gaussian Mixture Models (GMM)
  - The traditional best model for text-independent speaker recognition
- Support Vector Machines (SVM)
  - More recent use of discriminative model

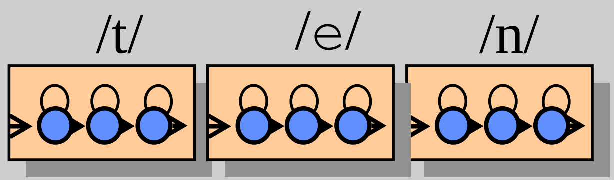
Form of GMM/HMM depends on application

Fixed Phrase → Word/phrase models

"Open sesame"

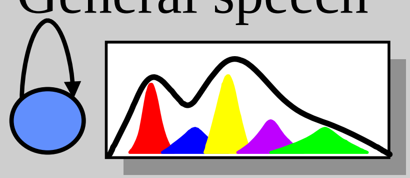


Prompted phrases/passwords → Phoneme models



Text-independent → single state HMM (GMM)

General speech

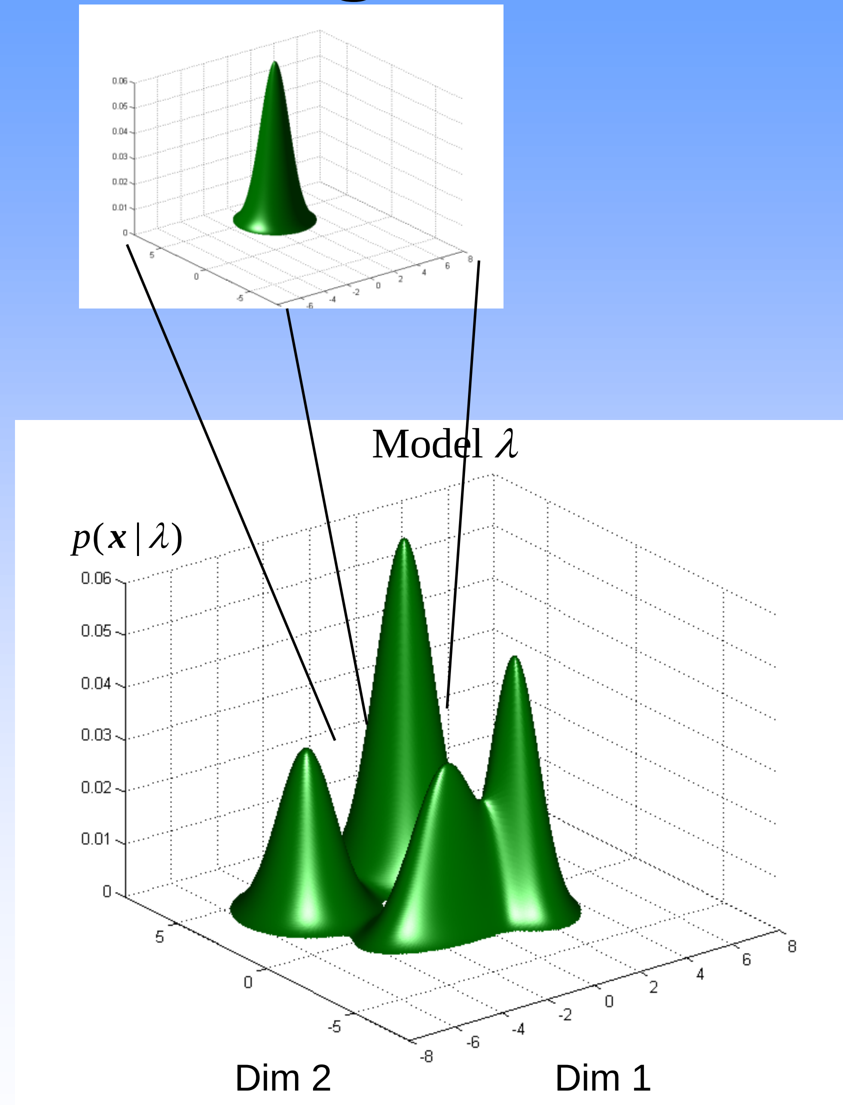


slide from Douglas Reynolds

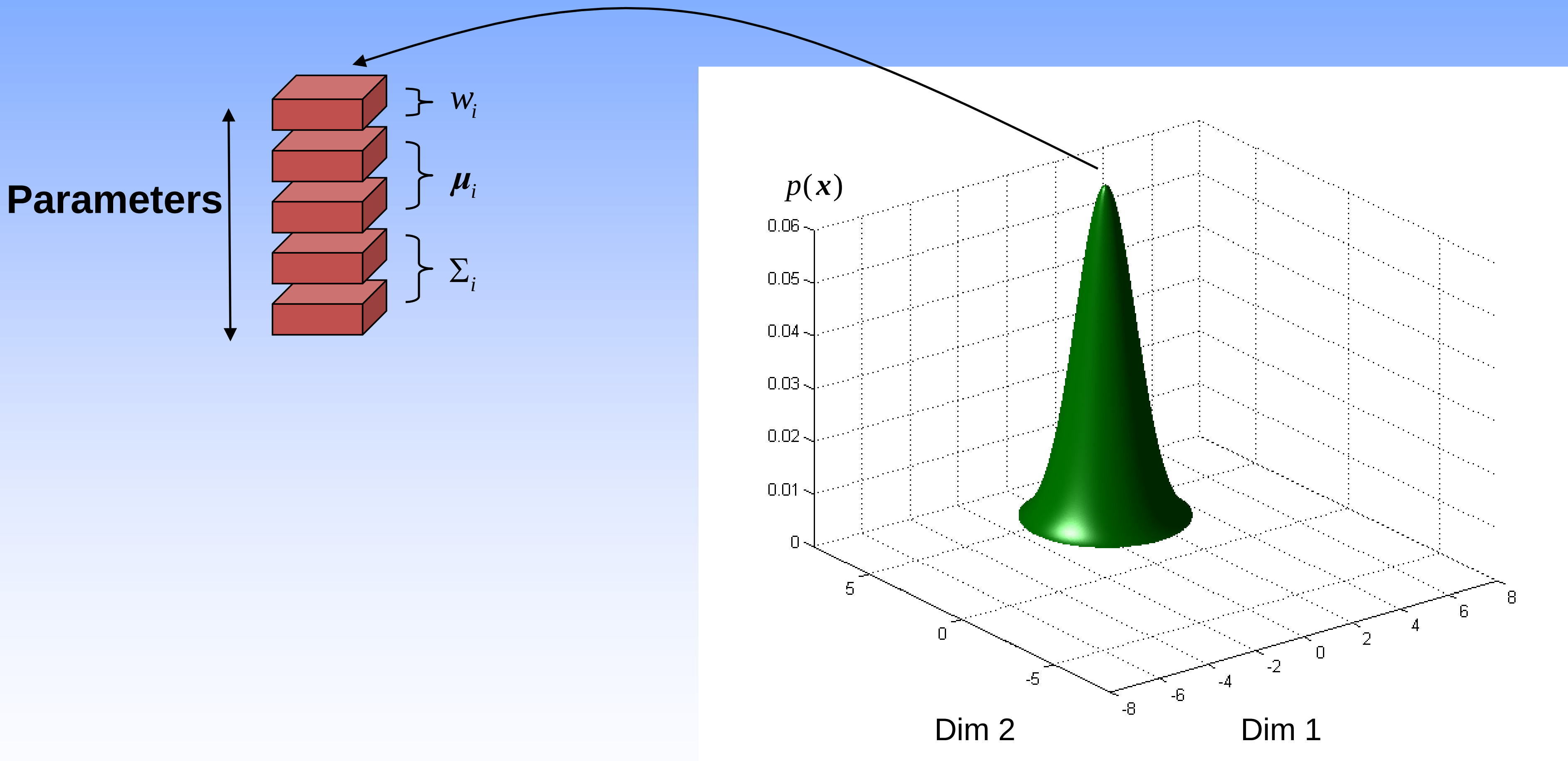
GMMs for speaker recognition

- A Gaussian mixture model (GMM) represents features as the weighted sum of multiple Gaussian distributions

- Each Gaussian state i has a
  – Mean $\boldsymbol{\mu}_i$
  – Covariance Σ_i
  – Weight w_i

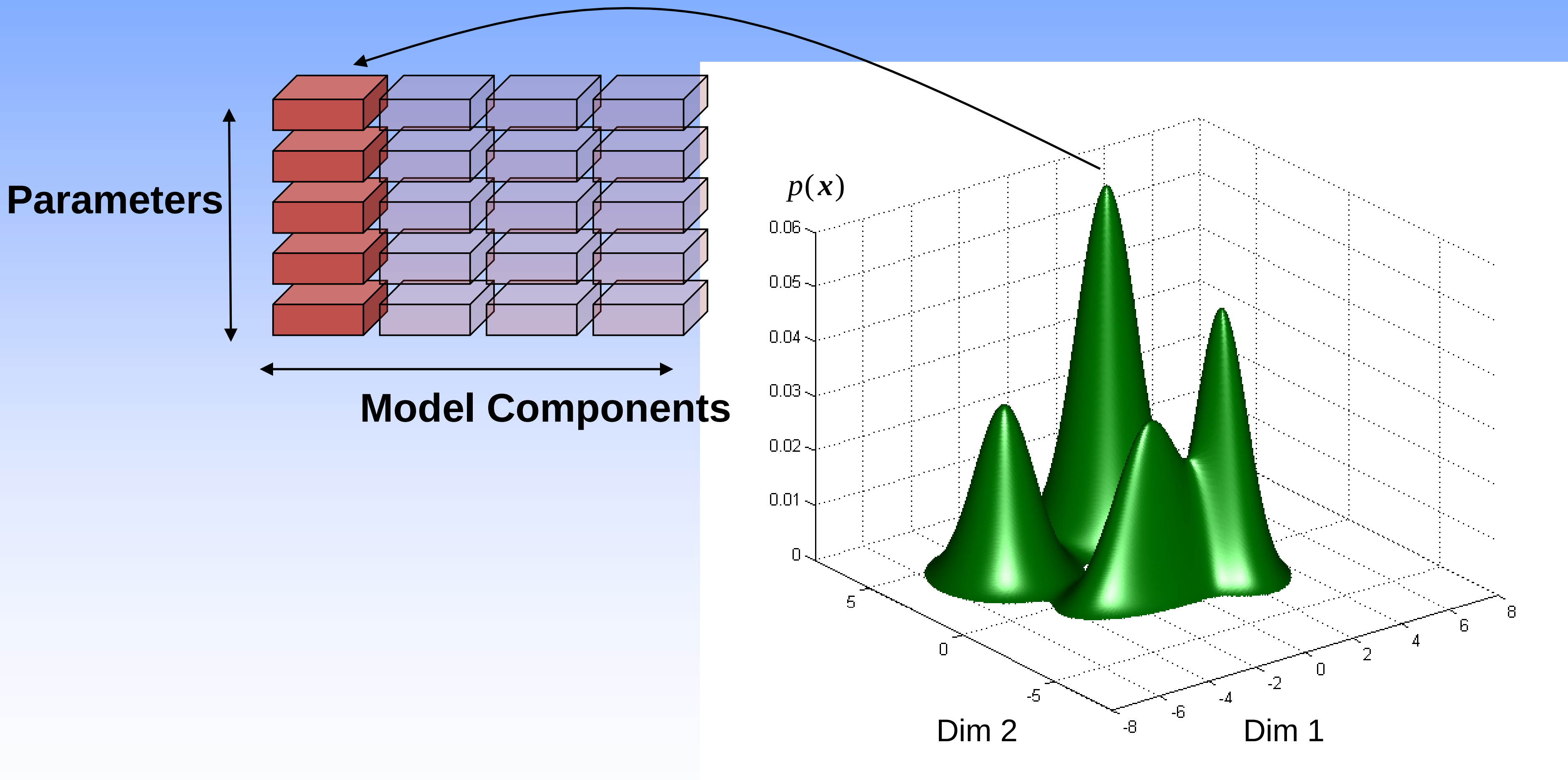


Recognition Systems
Gaussian Mixture Models



Nicolas Malyska, Sanjeev Mohindra, Karen Lauro, Douglas Reynolds, and Jeremy Kepner

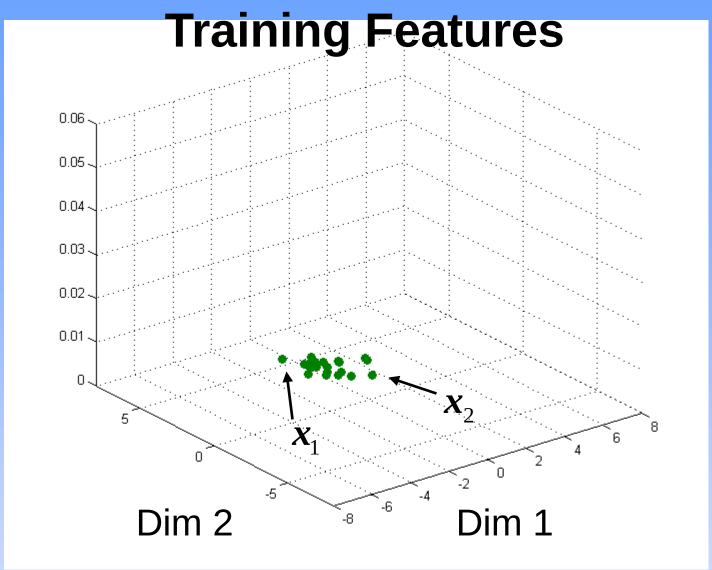
Recognition Systems
Gaussian Mixture Models

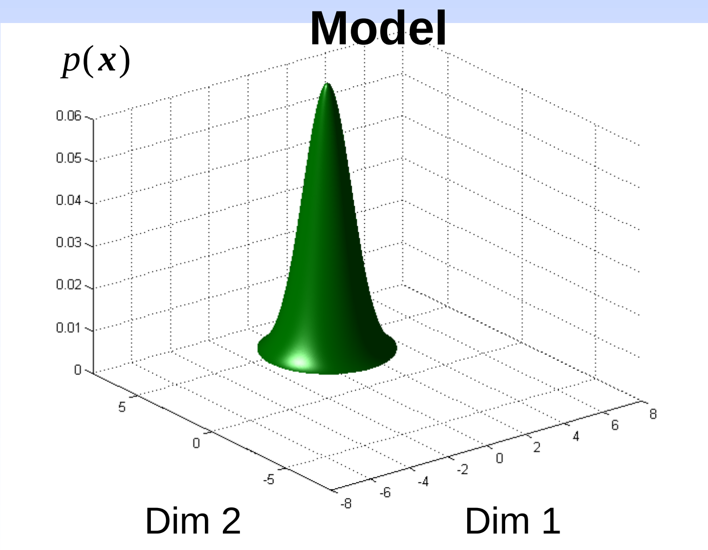


Nicolas Malyska, Sanjeev Mohindra, Karen Lauro, Douglas Reynolds, and Jeremy Kepner

GMM training

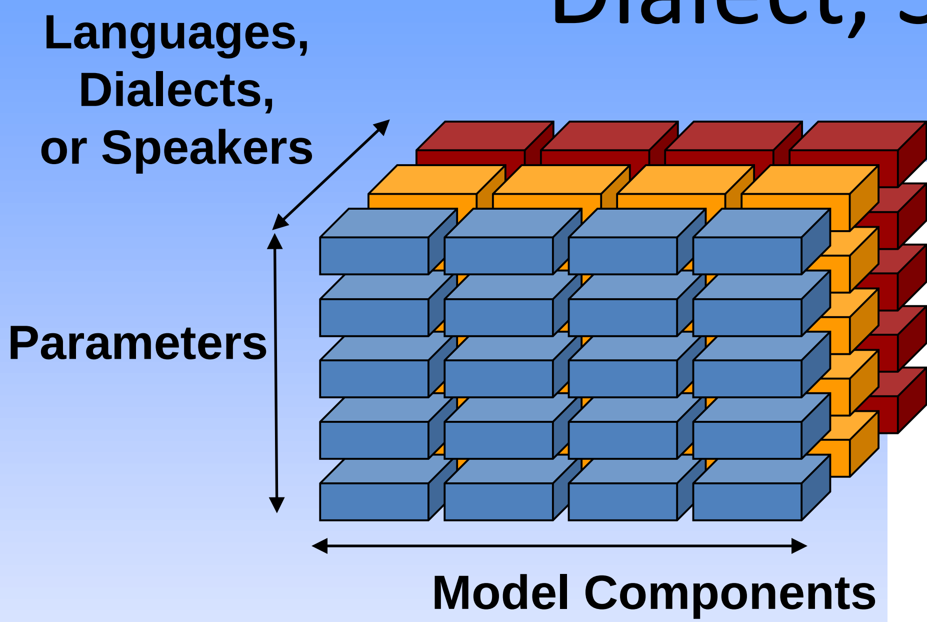
- During training, the system learns about the data it uses to make decisions
  - A set of features are collected from a speaker (or language or dialect)

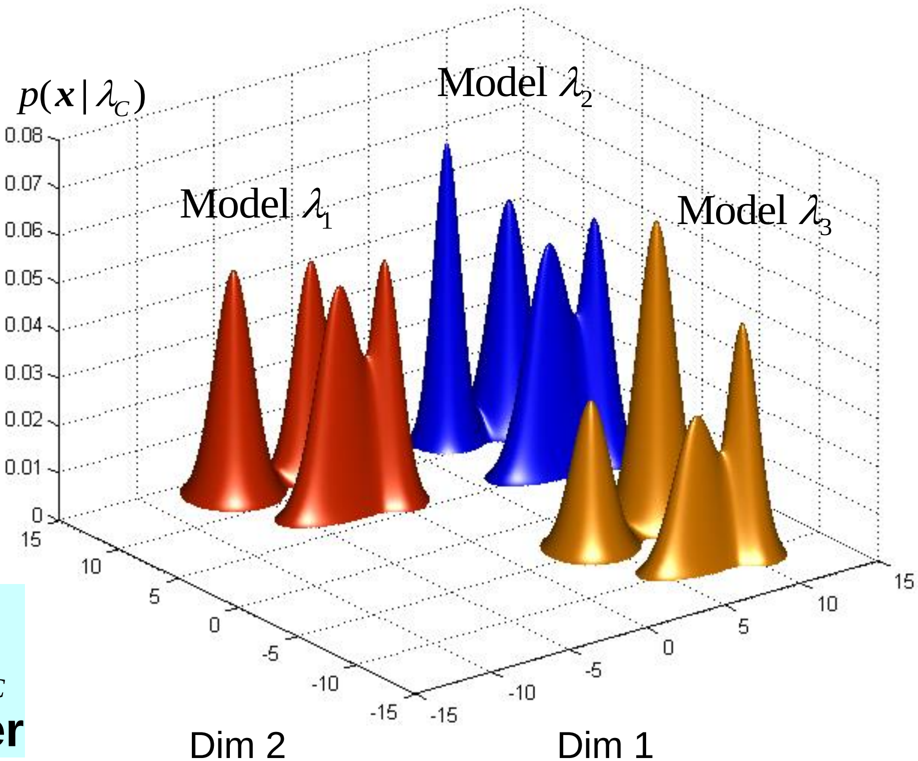




Nicolas Malyska, Sanjeev Mohindra, Karen Lauro, Douglas Reynolds, and Jeremy Kepner

Recognition Systems for Language, Dialect, Speaker ID

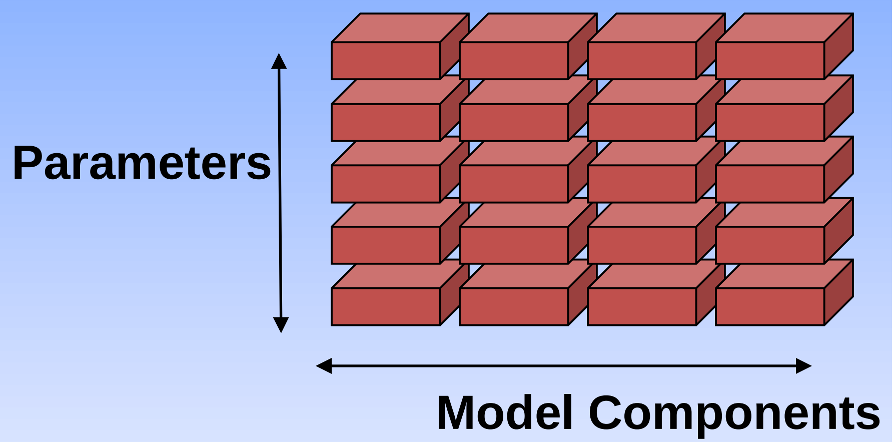


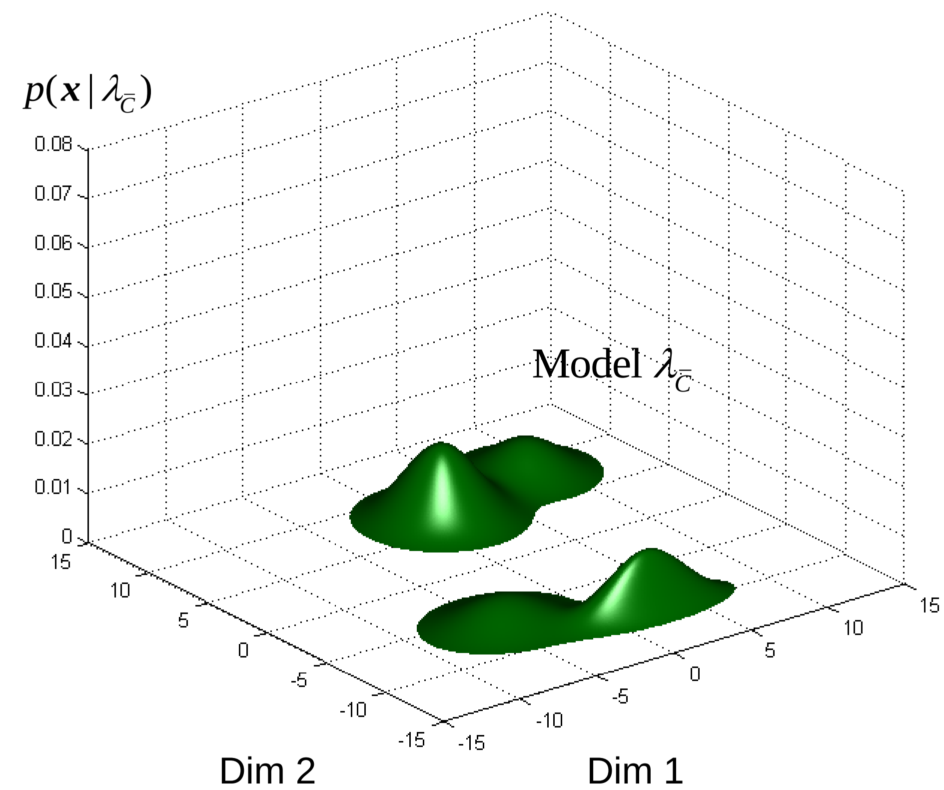


In LID, DID, and SID,
we train a set of *target models* λ_C
for each dialect, language, or speaker

Nicolas Malyska, Sanjeev Mohindra, Karen Lauro, Douglas Reynolds, and Jeremy Kepner

Recognition Systems
Universal Background Model





We also train a *universal background model* $\lambda_{\bar{C}}$ representing all speech

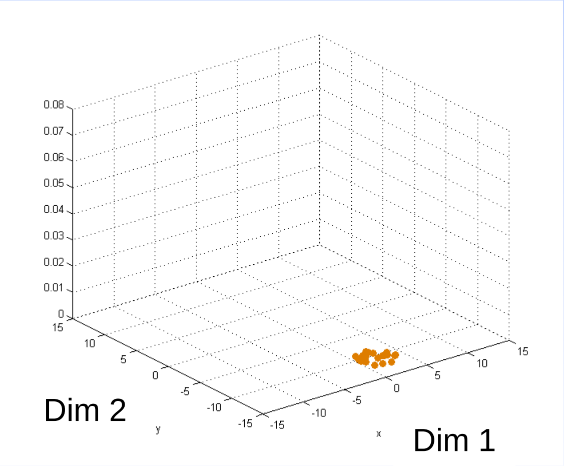
Nicolas Malyska, Sanjeev Mohindra, Karen Lauro, Douglas Reynolds, and Jeremy Kepner

Recognition Systems
Hypothesis Test

- Given a set of test observations, we perform a hypothesis test to determine whether a certain class produced it

$$H_0: \quad X_{test} \text{ is from the hypothesized class}$$
$$H_1: \quad X_{test} \text{ is not from the hypothesized class}$$

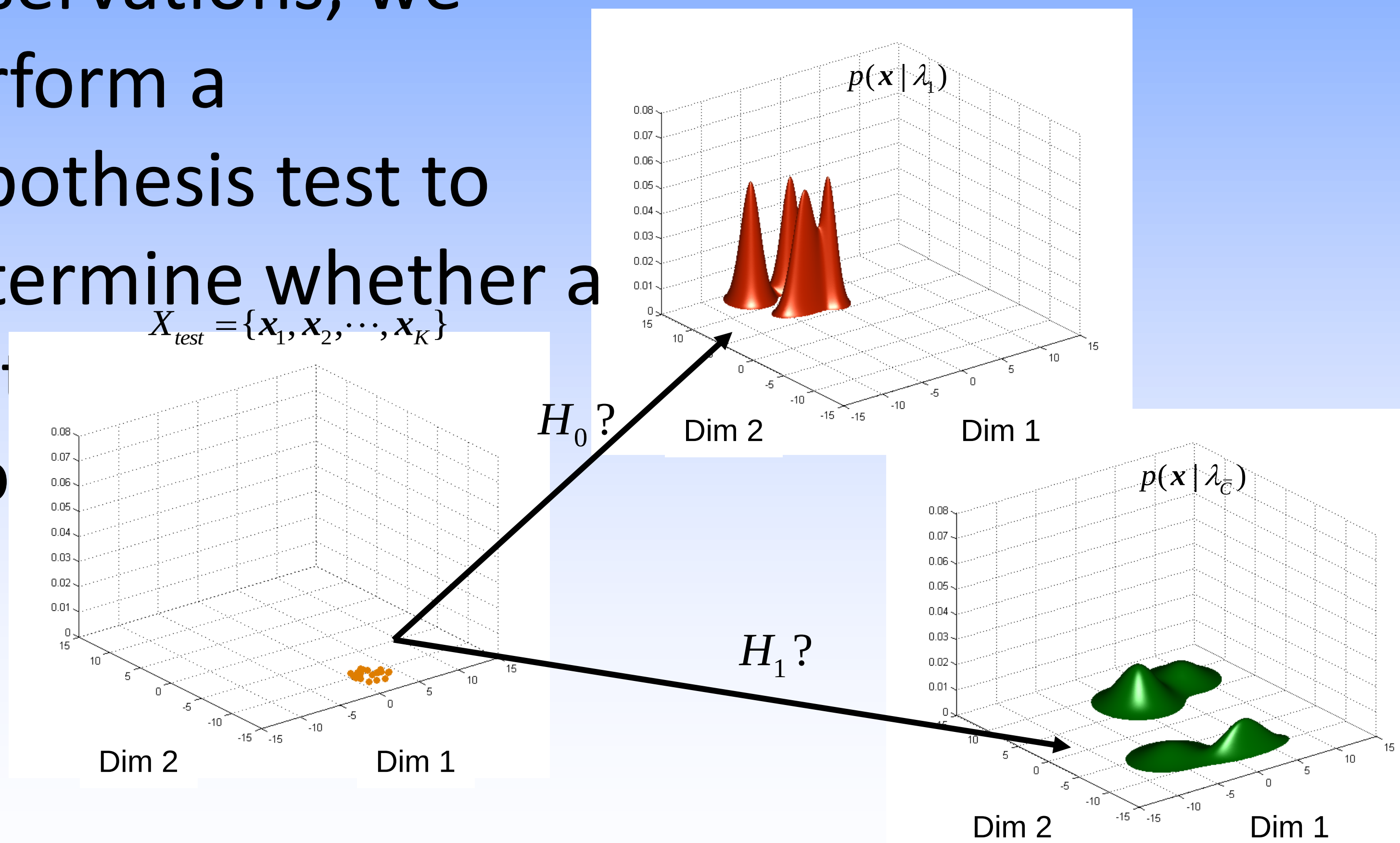
$$X_{test} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_K\}$$



Recognition Systems
Hypothesis Test

$$H_0: \quad X_{test} \text{ is from the hypothesized class}$$
$$H_1: \quad X_{test} \text{ is not from the hypothesized class}$$

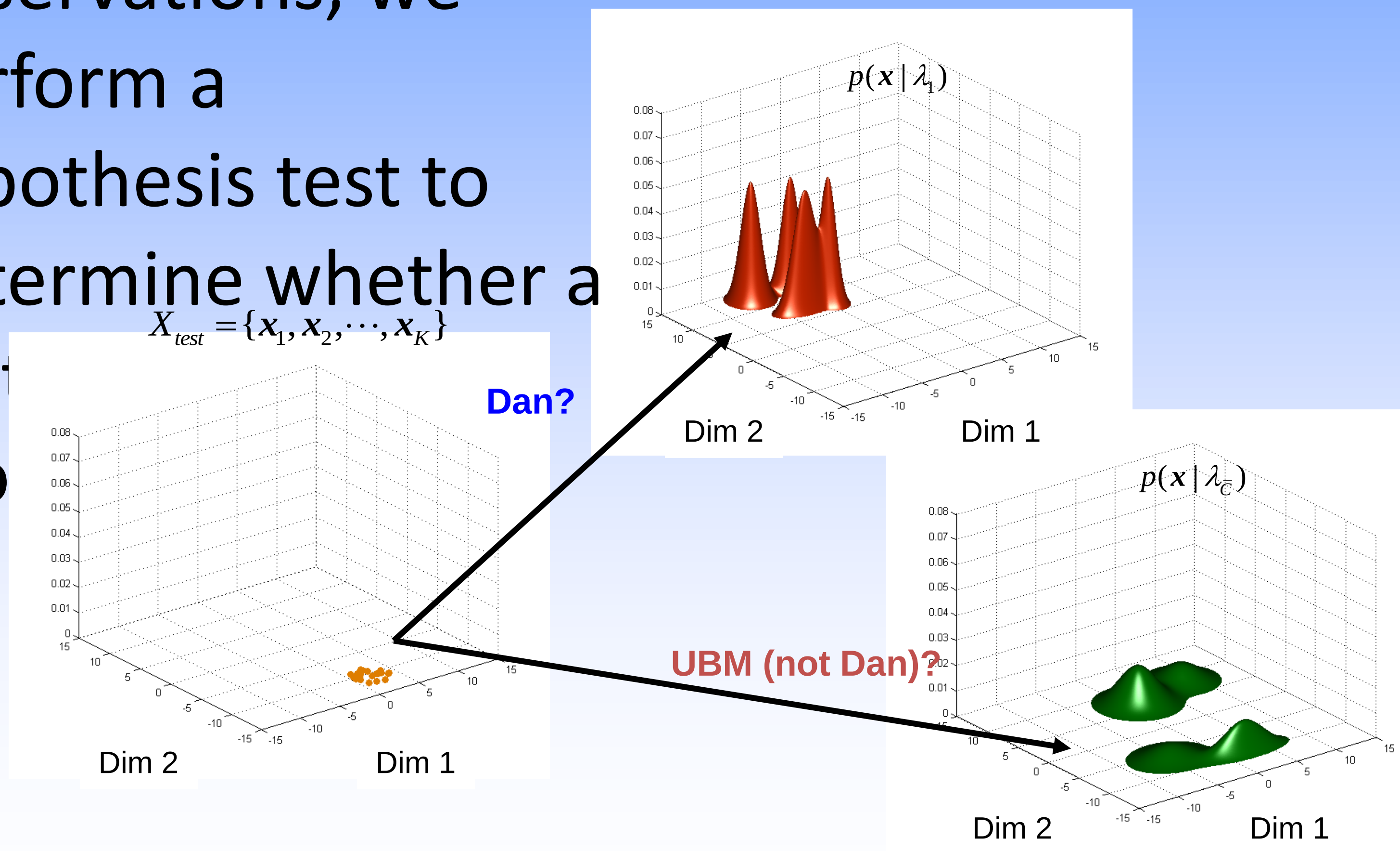
- Given a set of test observations, we perform a hypothesis test to determine whether a cer[...] pro[...]



Nicolas Malyska, Sanjeev Mohindra, Karen Lauro, Douglas Reynolds, and Jeremy Kepner

Recognition Systems
Hypothesis Test

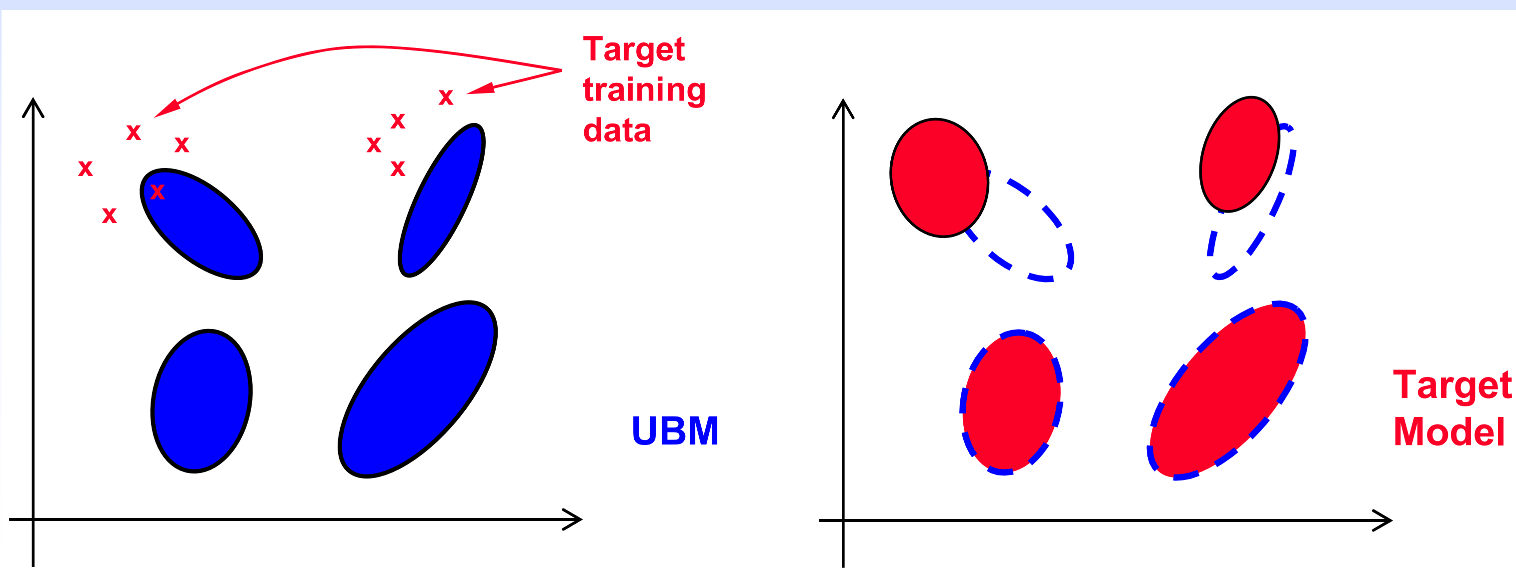
- Given a set of test observations, we perform a hypothesis test to determine whether a cer
pro



More details on GMMs

- Instead of training speaker model on only speaker data
- Adapt the UBM to that speaker
  - takes advantage of all the data
  - MAP adaptation: new mean of each Gaussian is a weighted mix of the UBM and the speaker

a:

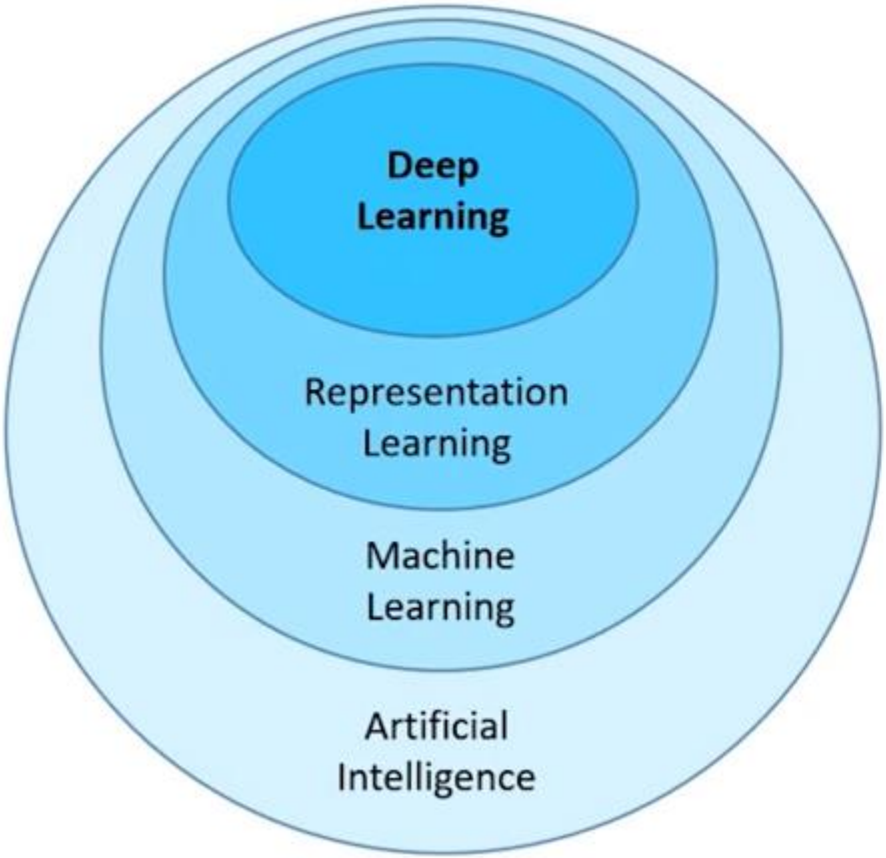


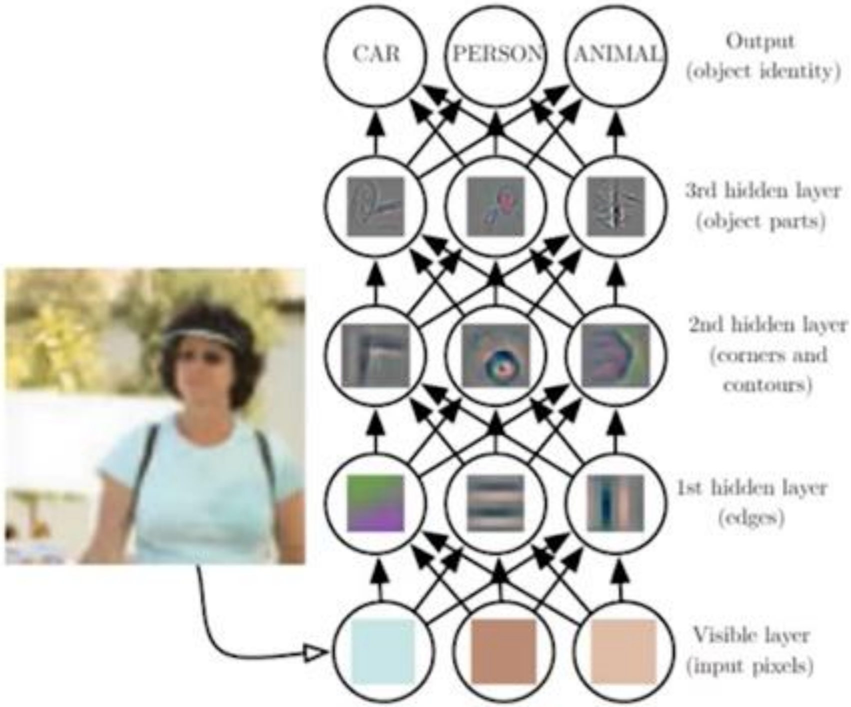
Gaussian mixture models

- Features are normal MFCC
  - can use more dimensions (20 + deltas)
- UBM background model: 512–2048 mixtures
- Speaker’s GMM: 64–256 mixtures
- Often combined with other classifiers in mixture-of-experts

Deep Learning is **Representation Learning**

(aka Feature Learning)





MIT Deep Learning Basics: Introduction and Overview

https://www.youtube.com/watch?v=O5xeyoRL95U