Course 495: Advanced Statistical Machine Learning/Pattern Recognition

- Lectures: Stefanos Zafeiriou

- Goal (Lectures): To present modern statistical machine learning/pattern recognition algorithms. The course focuses on statistical latent variable models (continuous & discrete).

- Goal (Tutorials): To provide the students the necessary mathematical tools for deeply understanding the models.

- Main Material: Pattern Recognition & Machine Learning by C. Bishop
  Chapters 1,2,12,13,8,9

- More materials in the website:
  http://ibug.doc.ic.ac.uk/courses/advanced-statistical-machine-learning-495/

- Email of the course: course495imperial@gmail.com

Statistical Machine Learning

The two main concepts are : **machine learning** and **statistics**

Machine Learning: A branch of Artificial Intelligence (A.I.) that focuses on designing, developing and studying the properties of algorithms that **learn** from data.

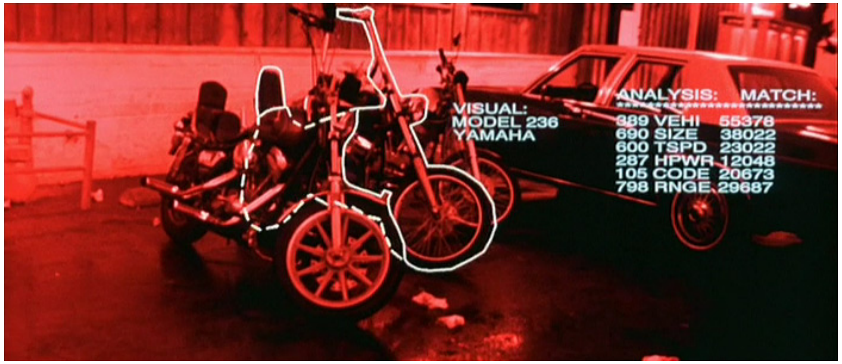
Statistics: A branch of applied mathematics that study collection, organization, analysis, interpretation, presentation and visualization of data and modelling their **randomness** and **uncertainty** using probability theory, as well as linear algebra and analysis.

Learn:

- Learning is at the core of the problem of **Intelligence**. Models of learning are used for understanding the function of the brain. Learning is used to develop modern **intelligent machines**.
- In many disciplines learning is now one of the main lines of research, i.e. signal/speech/ (medical) image processing, computer vision, control/robotics, natural language processing, bioinfomatics etc.
- It starts to dominate other domains such as software engineering, security sciences, finance etc.
- Major players invest huge amount of money in modern learning systems (e.g., Deep Learning by Google and Facebook)
- Next 25 years will be the age of machine learning

Machine Learning in movies

Terminator 2



Machine Learning in movies

Hal 9000

Movie: 2001: A Space Odyssey

Director: Stanley Kubrick

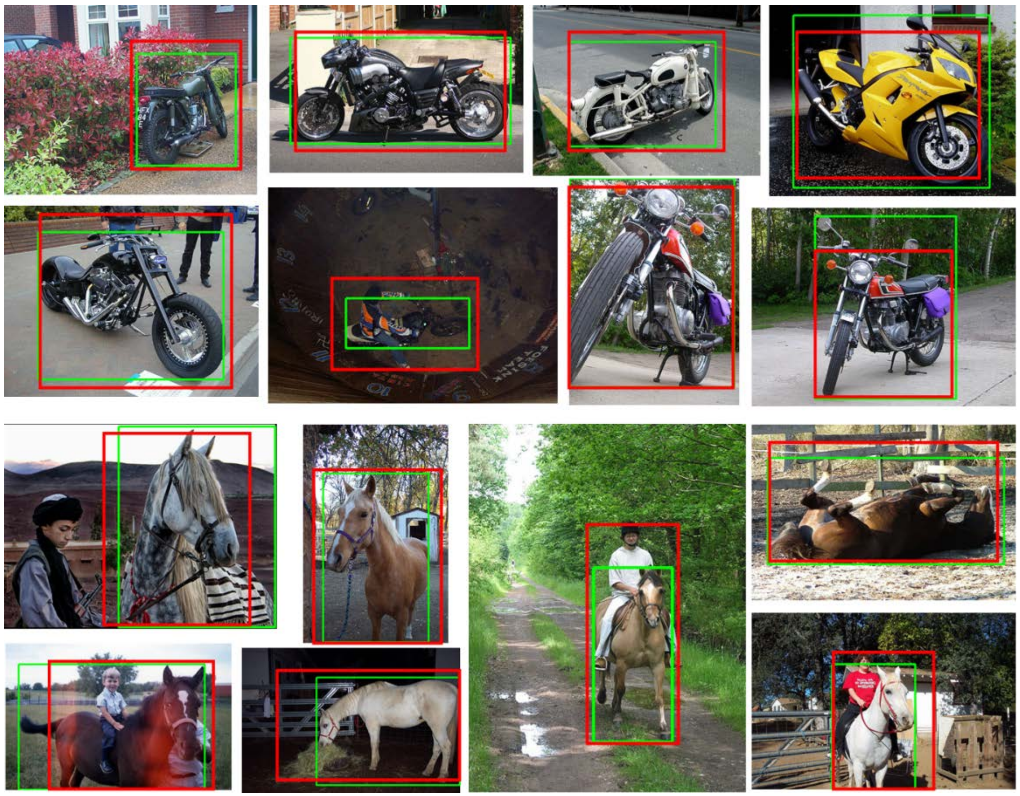
Statistical Machine Learning

Some applications of modern learning algorithms

- Face detection
- Object/Face tracking
- Biometrics
- Speech/Gesture recognition
- Image segmentation
- Finance
- Bioinformatics

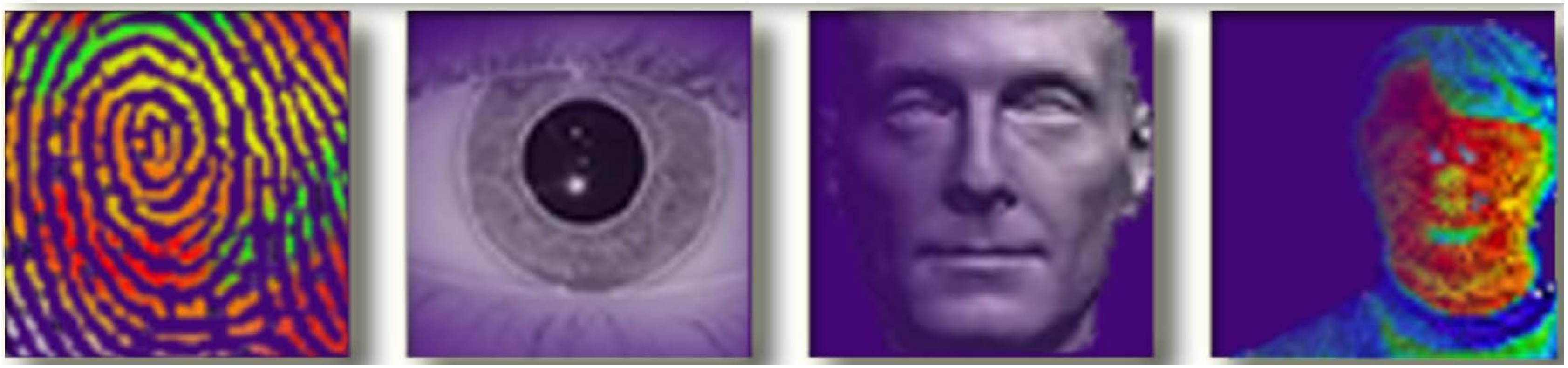
Applications

Object & face detection (modern cameras)



Applications

Face/Iris/Fingerprint recognition (biometrics)



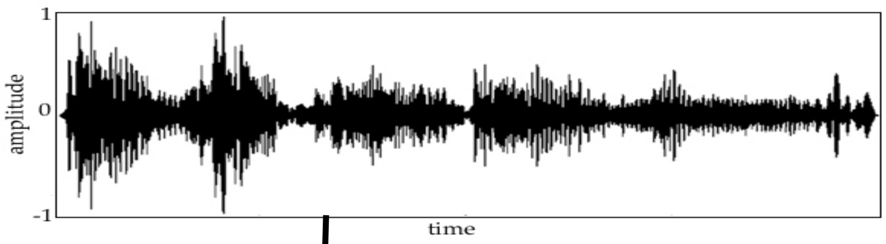
Applications

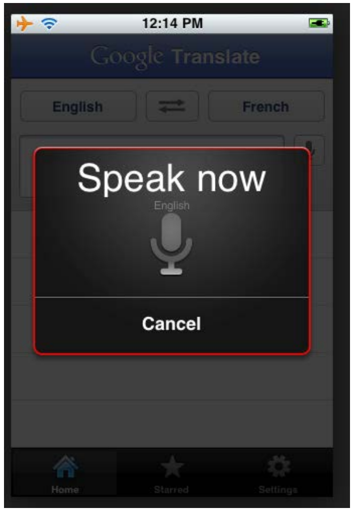
Object-target tracking

Applications

Speech Recognition (voice Google search)

Waveform

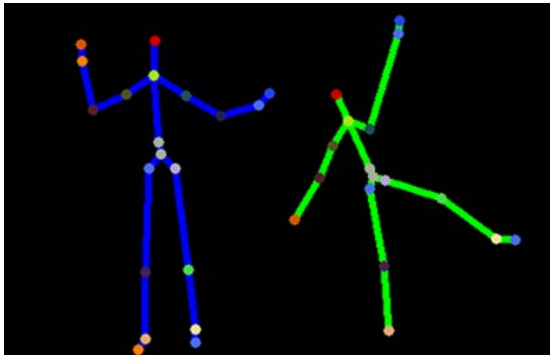




Hello world

Applications

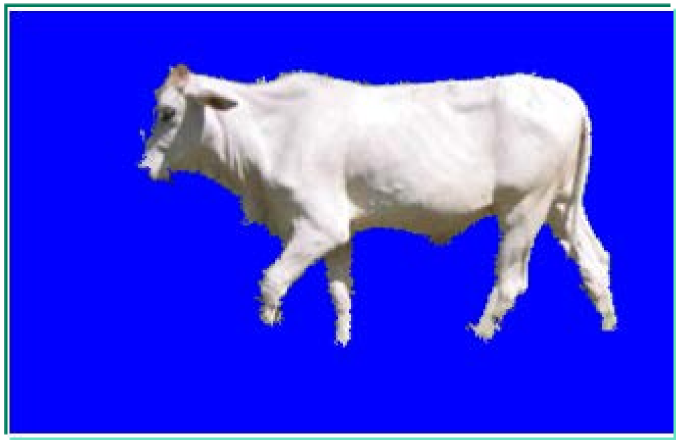
Gesture recognition (Kinect games)



Gestures

Applications

Image Segmentation



Applications

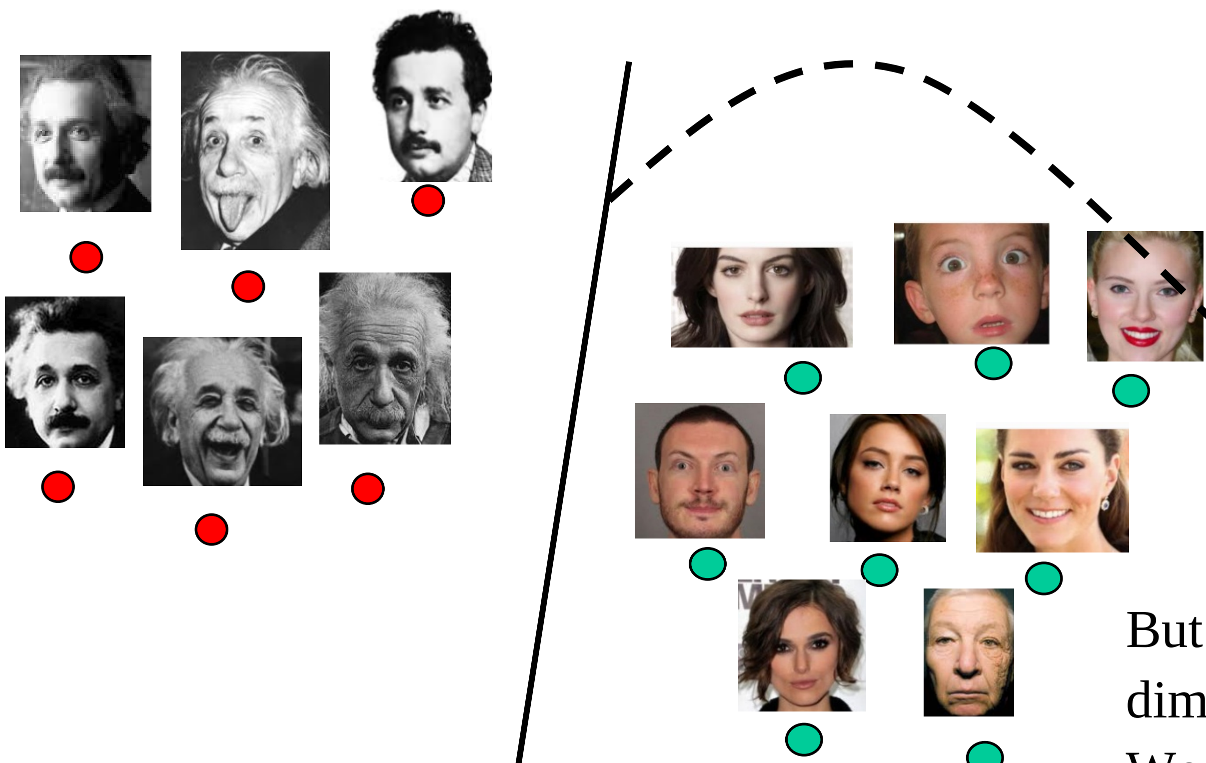
Biological data

Discover genes associated to a disease

what are you thinking?

What does the machine learn in each application?

Face Recognition: learn a classifier, i.e. a function (design of classifiers is covered in 424: Neural Computation)

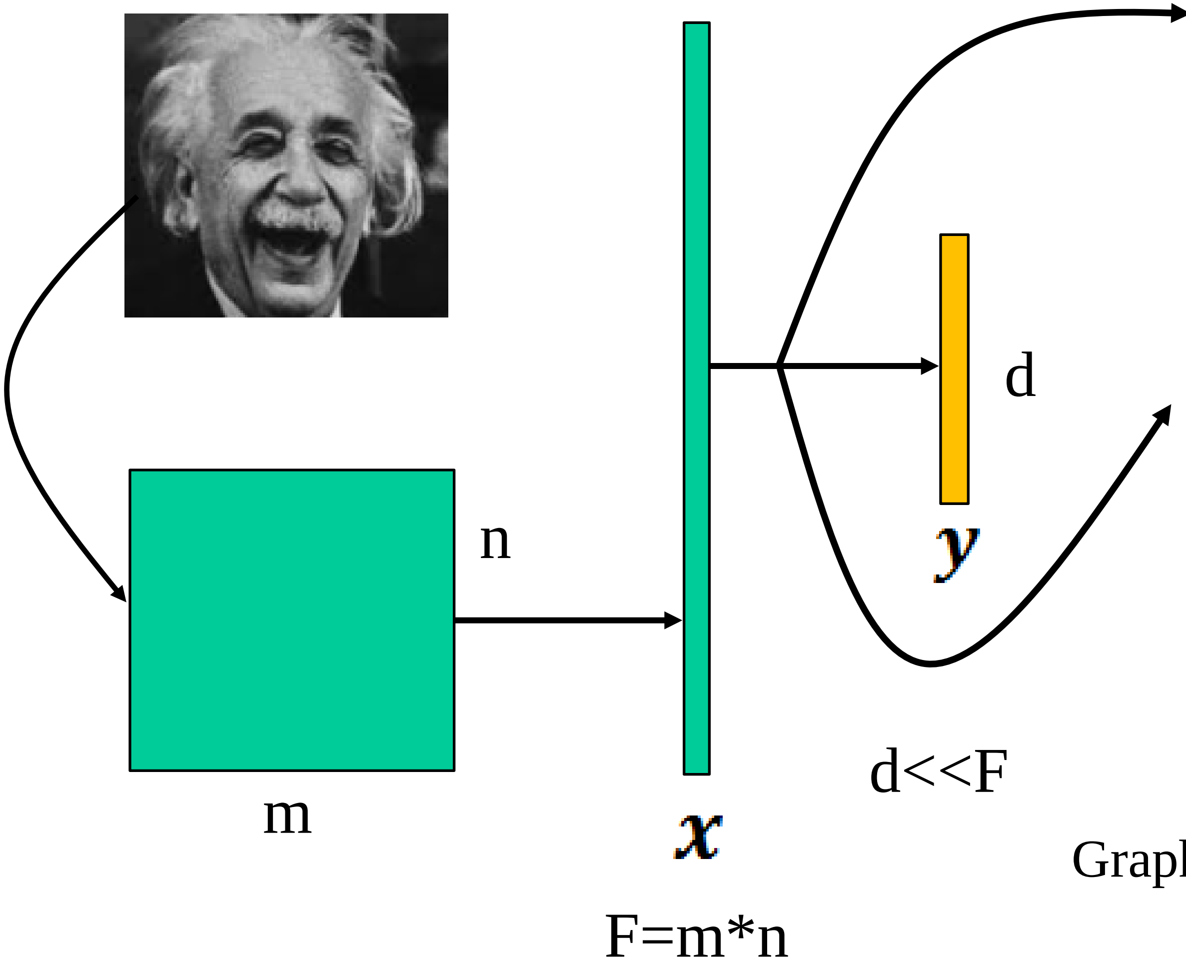


Model: $y_i = \boldsymbol{w}^T \boldsymbol{x_i} + \boldsymbol{b}$

Parameters: $\theta = \{\mathbf{w}, b\}$

But images are very high dimensional objects!!!
We need to reduce them!!

Latent Variable Models



Model (Deterministic):

$$\boldsymbol{y} = \boldsymbol{W}^T \boldsymbol{x}$$

Parameters: $\theta = \{\boldsymbol{W}\}$

Model (probabilistic):

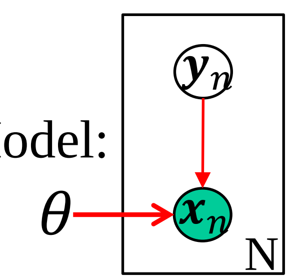
$$\boldsymbol{x} = \boldsymbol{W}\boldsymbol{y} + \boldsymbol{\mu} + \boldsymbol{e}$$

$$\boldsymbol{e} \sim N(\boldsymbol{e}|\boldsymbol{0}, \sigma^2 \boldsymbol{I})$$

$$\boldsymbol{y} \sim N(\boldsymbol{y}|\boldsymbol{0}, \boldsymbol{I})$$

Parameters: $\theta = \{\boldsymbol{W}, \boldsymbol{\mu}, \sigma^2\}$

Graphical Model:



Latent Variable Models

- Deterministic model:
  - ✓ There is no randomness (uncertainty) associated with the model.
  - ✓ We can compute the actual values of the latent space.
- Probabilistic model:
  - ✓ We assign probability distributions to the latent variables and model their dependencies.
  - ✓ We can compute only statistics of the latent variables.
  - ✓ More flexible than deterministic models.
- Generative Probabilistic models:
  - ✓ Model observations drawn from a probability density function (pdf) (i.e., model the way data are "generated")

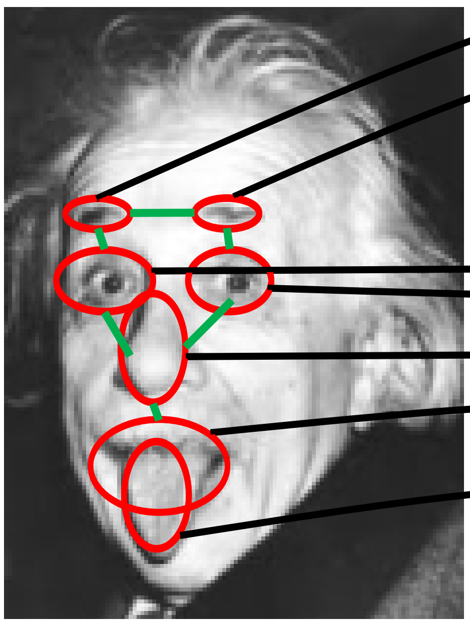
Latent Variable Models

•Generative Probabilistic models:

✓ Model the complete (joint) likelihood of both the data and latent structure

•But what is the latent (hidden) structure:

Intuitive.



{left eyebrow}
{right eyebrow}

{left eye}
{right eye}

{nose}

{mouth}
{tongue}

Latent (hidden) structure

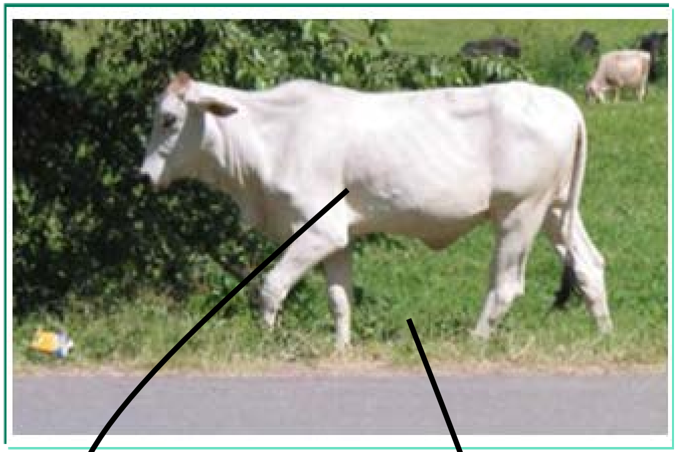
Latent Variable Models

Speech:

Word: need

Phonemes: n iy d

Image:



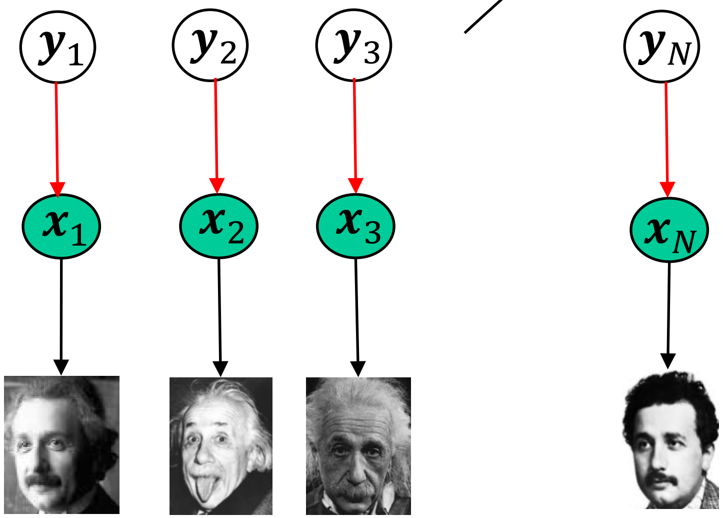
Object

Background

Latent structure

Latent Variable Models (Static)

General Concept:



Share a common linear structure

$$\boldsymbol{x} = \boldsymbol{W}\boldsymbol{y} + \boldsymbol{\mu} + \boldsymbol{e}$$

$$\boldsymbol{e} \sim N(\boldsymbol{e}|\boldsymbol{0}, \sigma^2\boldsymbol{I})$$

$$\boldsymbol{y} \sim N(\boldsymbol{y}|\boldsymbol{0}, \boldsymbol{I})$$

We want to find the parameters:

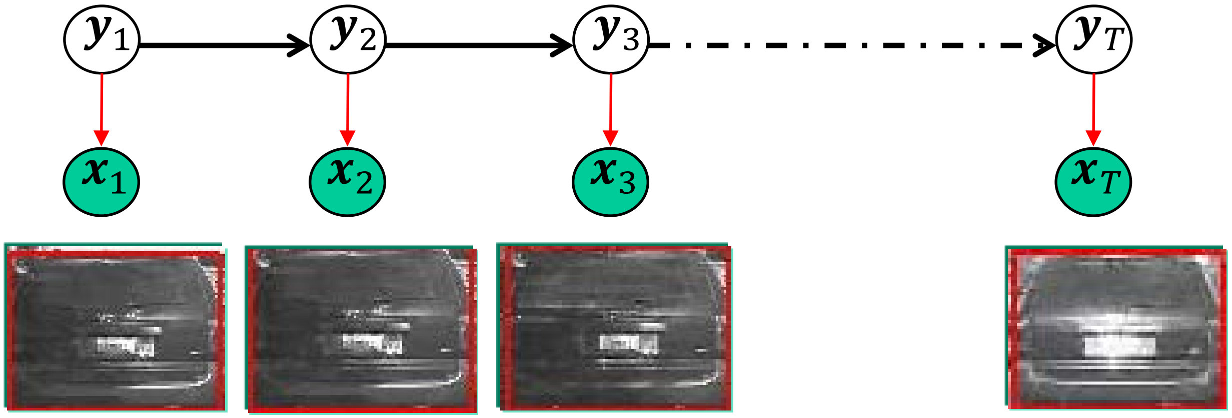
$$\theta = \{\boldsymbol{W}, \boldsymbol{\mu}, \sigma^2\}$$

Joint likelihood maximization:

$$\mathrm{p}(\boldsymbol{x}_1, .., \boldsymbol{x}_N, \boldsymbol{y}_1, .., \boldsymbol{y}_N|\theta) = \prod_{i=1}^{N} p(\boldsymbol{x}_i|\boldsymbol{y}_i, \boldsymbol{W}, \boldsymbol{\mu}, \sigma) \prod_{i=1}^{\mathrm{N}} \boldsymbol{p}(\boldsymbol{y}_i)$$

Latent Variable Models (Dynamic, Continuous)

Latent Variable Models (Dynamic, Continuous)



Generative Model

$$\boldsymbol{x}_n = \mathbf{W}\boldsymbol{y}_n + \boldsymbol{e}_n$$

$$\boldsymbol{y}_1 = \boldsymbol{\mu}_0 + \boldsymbol{u}$$

$$\boldsymbol{y}_n = \boldsymbol{A}\boldsymbol{y}_{n-1} + \boldsymbol{v}_n$$

Noise distribution

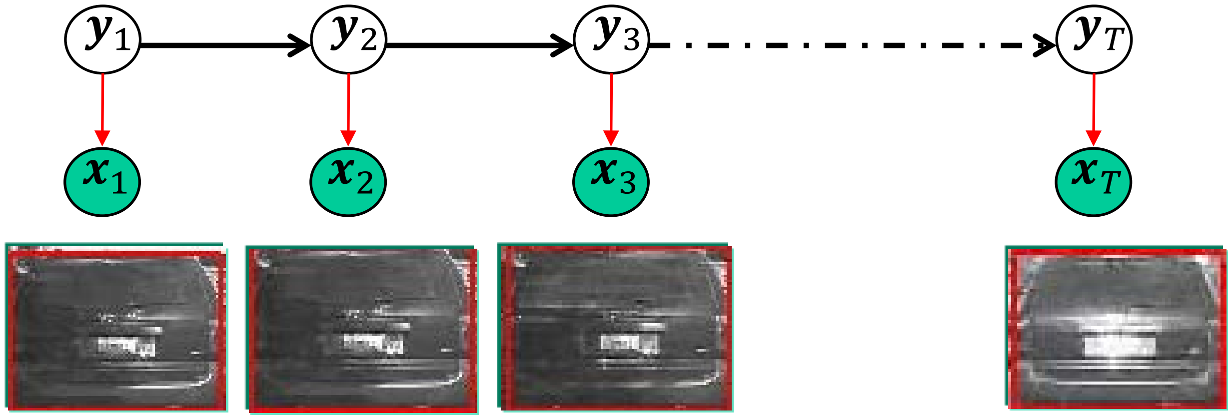
$$\mathbf{e} \sim N(\boldsymbol{e}|\mathbf{0}, \mathbf{\Sigma})$$

$$\boldsymbol{u} \sim N(\boldsymbol{u}|\mathbf{0}, \boldsymbol{P}_0)$$

$$\boldsymbol{v} \sim N(\boldsymbol{v}|\mathbf{0}, \mathbf{\Gamma})$$

Parameters: $\theta = \{\boldsymbol{W}, \boldsymbol{A}, \boldsymbol{\mu}_0, \mathbf{\Sigma}, \mathbf{\Gamma}, \boldsymbol{P}_0\}$

Latent Variable Models (Dynamic, Continuous)



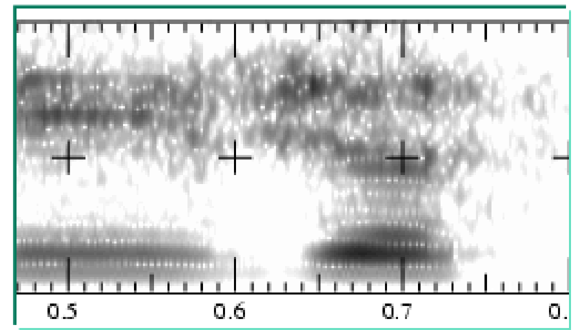
Markov Property: $p(\boldsymbol{y}_i, |\boldsymbol{y}_1, .., \boldsymbol{y}_{i-1}) = p(\boldsymbol{y}_i|\boldsymbol{y}_{i-1})$

Joint likelihood:

$$p(\boldsymbol{x}_1, .., \boldsymbol{x}_{T,}\boldsymbol{y}_1, .., \boldsymbol{y}_T)$$

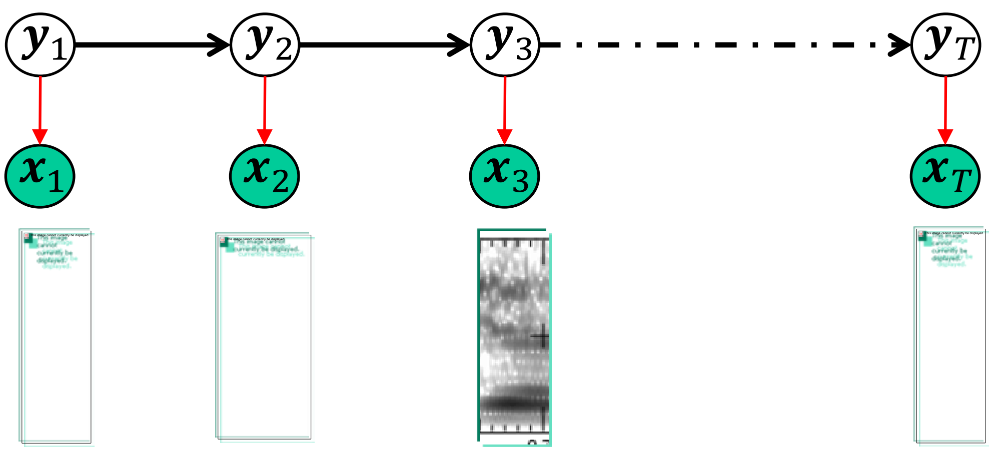
$$= \prod_{i=1}^{N} p(\boldsymbol{x}_i|\boldsymbol{y}_i, \boldsymbol{W}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma})p(\boldsymbol{y}_1|\boldsymbol{\mu}_0, \boldsymbol{P}_0)\prod_{\iota=2}^{\mathrm{N}} p(\boldsymbol{y}_i|\boldsymbol{y}_{i-1}, \mathbf{A}, \boldsymbol{\Gamma})$$

Latent Variable Models (Dynamic, Discrete)



Word: need

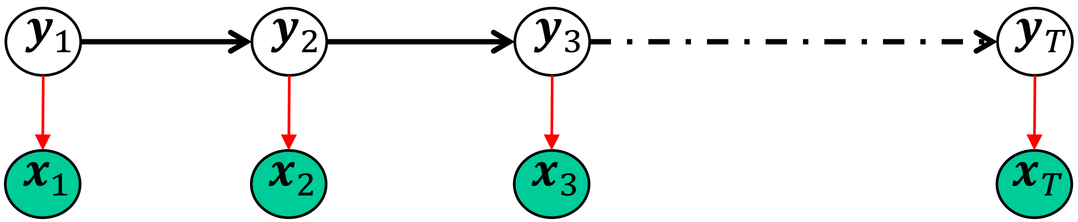
Phonemes: n iy d



Latent structure takes discrete values:

$y_t \in \{\text{start,n,iy,d,end}\}$

Latent Variable Models (Dynamic, Discrete)



$$\boldsymbol{A} = [a_{ij}] = [p(\boldsymbol{y}_t|\boldsymbol{y}_{t-1})]$$

$$\boldsymbol{\pi} = [p(\boldsymbol{y}_1)]$$

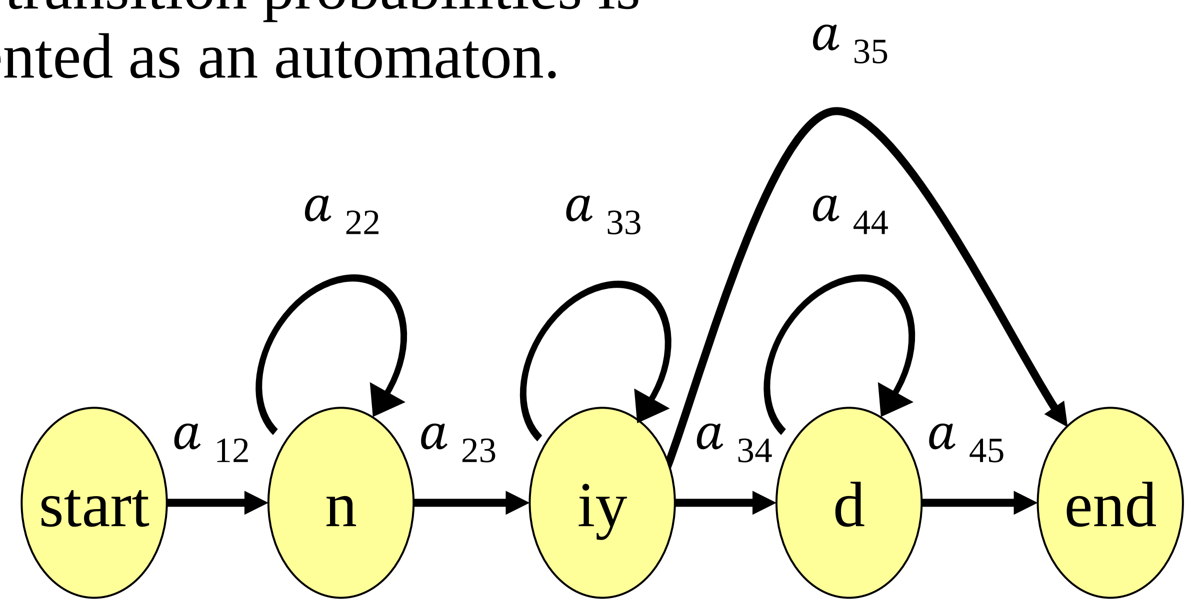
$\boldsymbol{y}_{t-1}$

$\boldsymbol{y}_t$

$p(\boldsymbol{y}_t\|\boldsymbol{y}_{t-1})$	s	n	iy	d	e
s	a_{11}	a_{12}	a_{13}	a_{14}	a_{15}
n	a_{21}	a_{22}	a_{23}	a_{24}	a_{25}
iy	a_{31}	a_{32}	a_{33}	a_{34}	a_{35}
d	a_{41}	a_{42}	a_{43}	a_{44}	a_{45}
e	a_{51}	a_{52}	a_{53}	a_{54}	a_{55}

Latent Variable Models (Dynamic, Discrete)

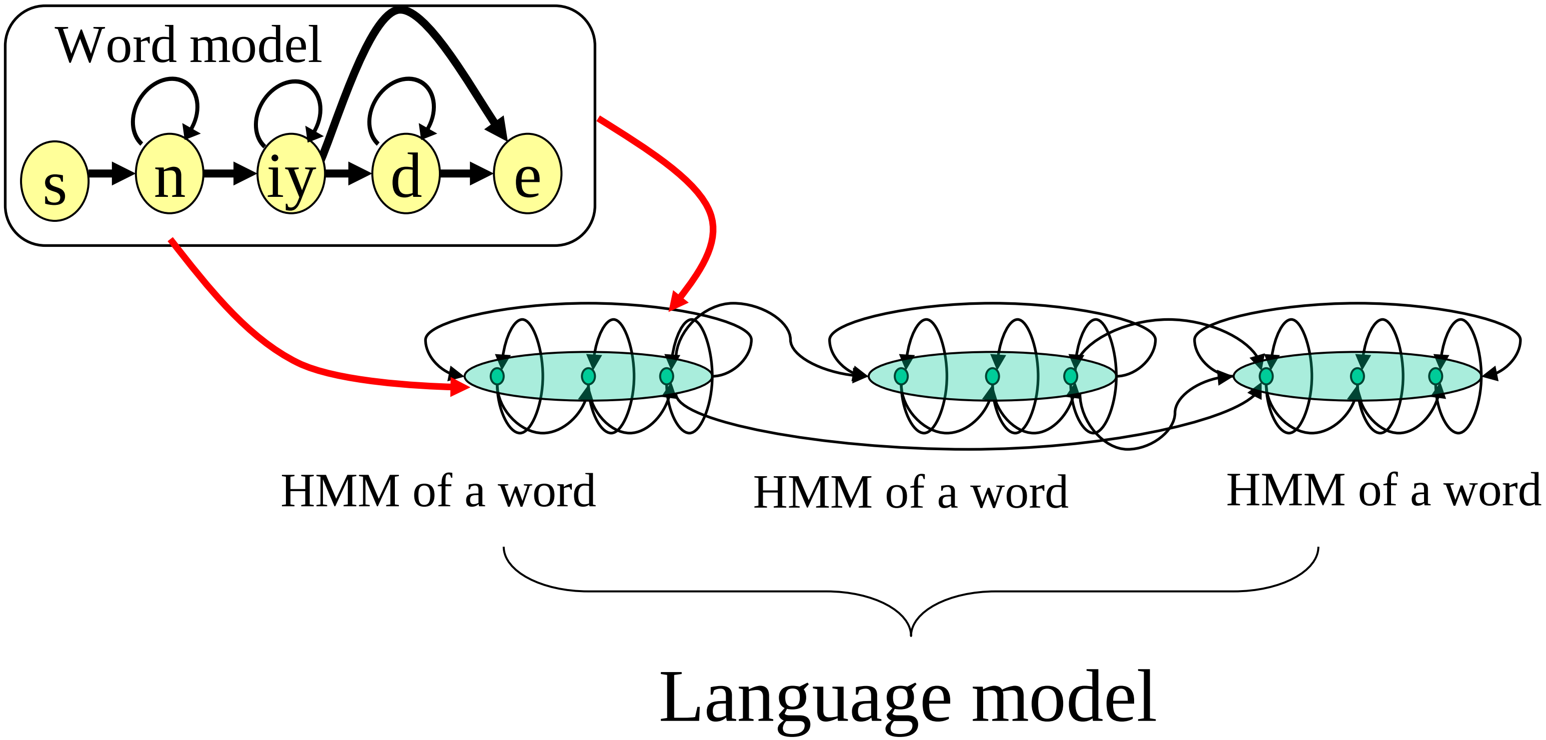
Matrix of transition probabilities is represented as an automaton.



Data generation: e.g. if the latent variable is $\boldsymbol{y}_t = \boldsymbol{b}$ $\quad \boldsymbol{x}_t \sim N(\boldsymbol{x}_t | \boldsymbol{m}_{\boldsymbol{b}}, \boldsymbol{S}_{\boldsymbol{b}})$

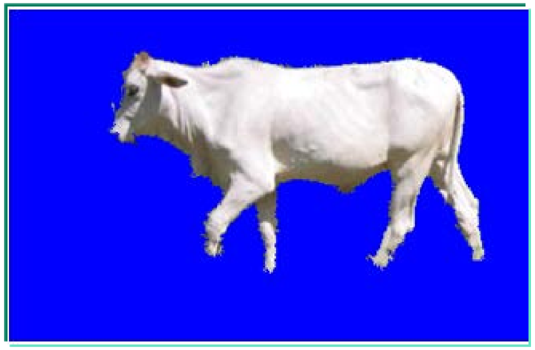
Parameters $\theta = \{\boldsymbol{A}, \{\boldsymbol{m}_{\boldsymbol{b}}, \boldsymbol{S}_{\boldsymbol{b}}\}_{\boldsymbol{b}}, \boldsymbol{\pi}\}$

Latent Variable Models (Dynamic, Discrete)



Latent Variable Models (Spatial)

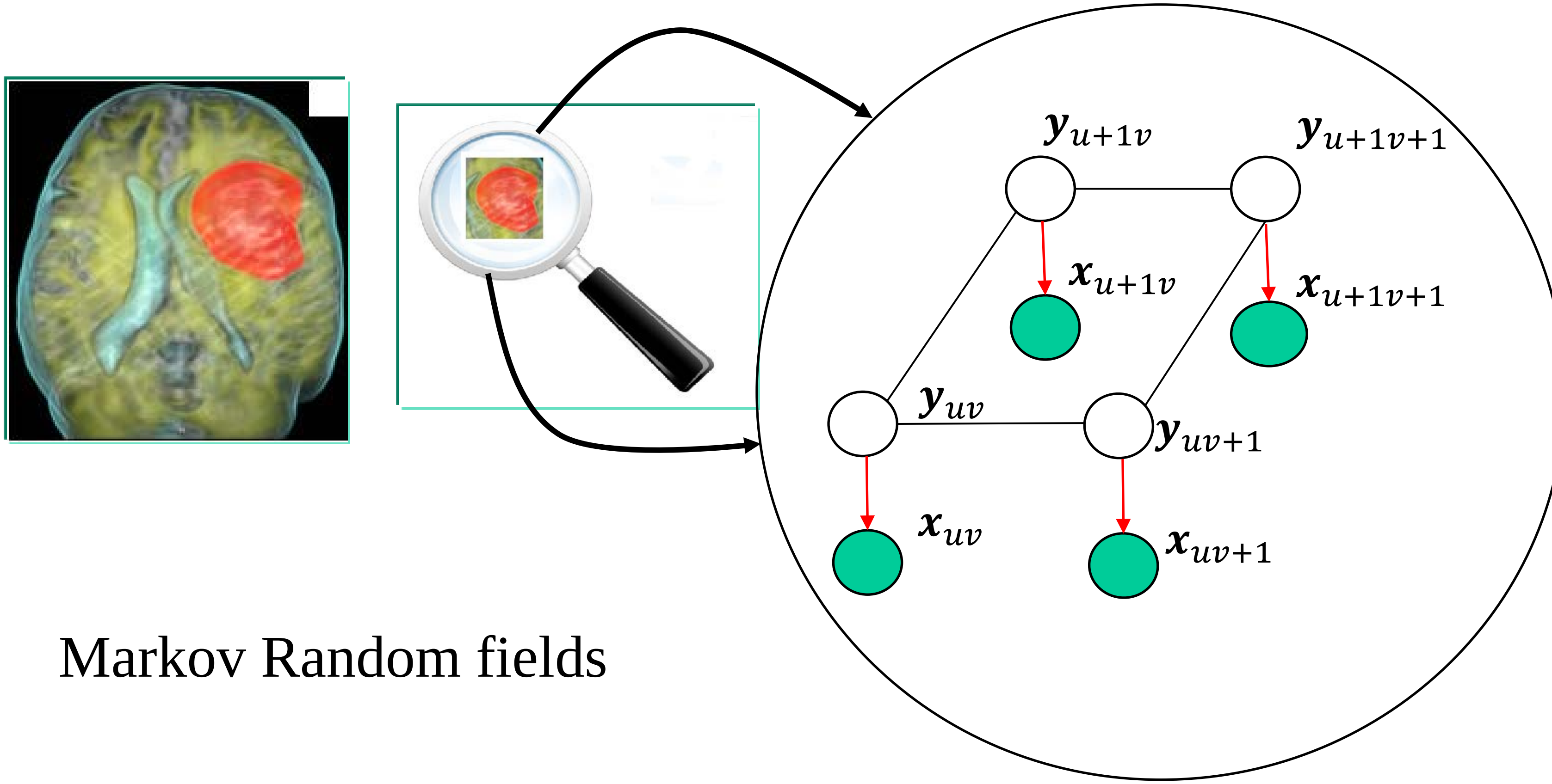
Image Segmentation



Brain tumour segmentation

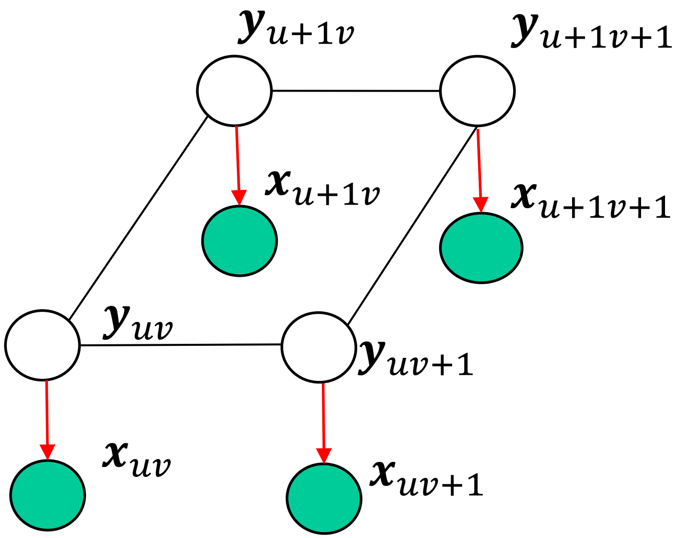
Latent Variable Models (Spatial)

Undirected spatial dependencies



Markov Random fields

Latent Variable Models (Spatial)



Markov Random fields

$$\mathrm{p}(\boldsymbol{y}_{11},..,\boldsymbol{y}_{nm}) = \frac{1}{Z}\prod_{C}\psi(\boldsymbol{y}_{C})$$

C is the maximal clique.

Potential function: $\psi(\boldsymbol{y}_C) = e^{(-E(\boldsymbol{y}_C))}$

Partition function: $Z = \sum_{y}\prod_{C}\psi(\boldsymbol{y}_C)$

Markov blanket:

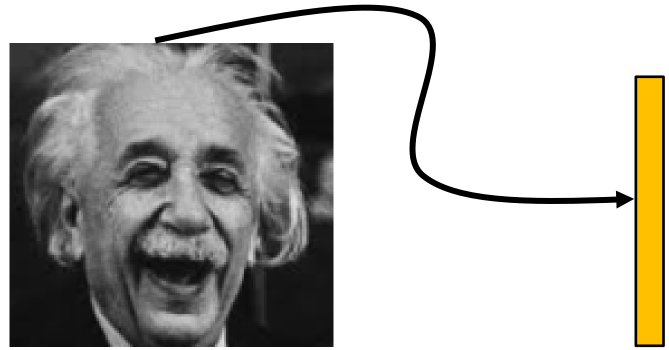
$$p(\boldsymbol{y}_{ul}, \boldsymbol{y}_{vk} | \boldsymbol{Y}_{/\,y_{ul},y_{vk}}) = p(\boldsymbol{y}_{ul} | \boldsymbol{Y}_{y_{ul}})\, p(\boldsymbol{y}_{vk} | \boldsymbol{Y}_{y_{vk}})$$

Complete likelihood:

$$p(\boldsymbol{X}, \boldsymbol{Y} | \theta) = \frac{1}{Z}\prod_{uv} p(\boldsymbol{x}_{uv} | \boldsymbol{y}_{uv}, \theta) \prod_{C} \psi(\boldsymbol{y}_C | \theta)$$

Summarize what we will study?

Deterministic Component Analysis (3 weeks)



$\boldsymbol{x}$ $\boldsymbol{y}$

Unsupervised approaches:

- Principal Component Analysis, Independent Component Analysis, Graph-based Component Analysis, Slow feature Analysis

Supervised approaches:

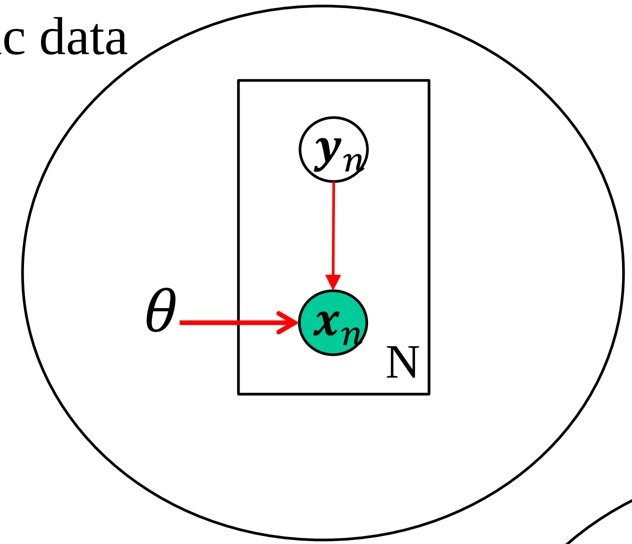
- Linear discriminant analysis

What we will learn?:

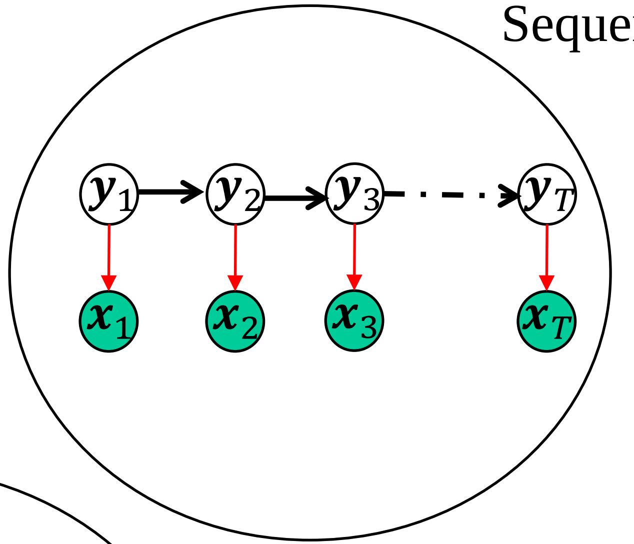
- *How to find the latent space directly* $\boldsymbol{y}$.
- *How to find the latent space via linear projections* $\boldsymbol{y} = \boldsymbol{W}^T \boldsymbol{x}$.

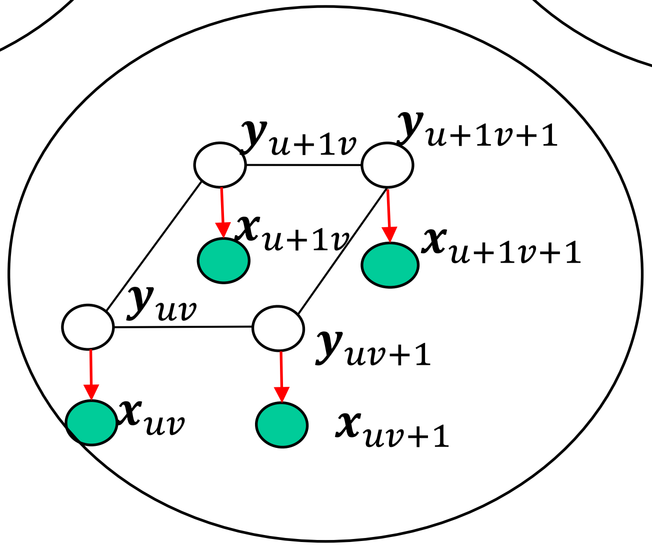
Summarize what we will study?

Static data



Sequential data

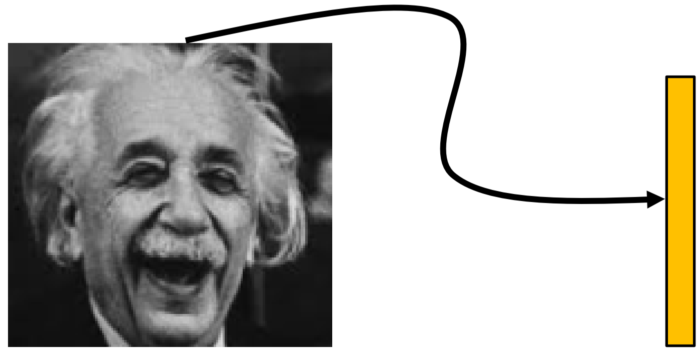




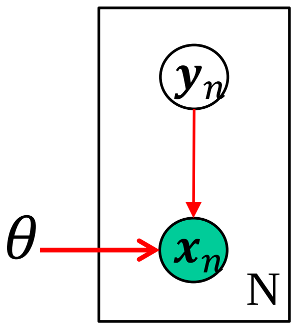
Spatial data

Summarize what we will study?

Probabilistic Principal component analysis



$\boldsymbol{x}$ $\qquad\qquad$ $\boldsymbol{E}[\boldsymbol{y}]$

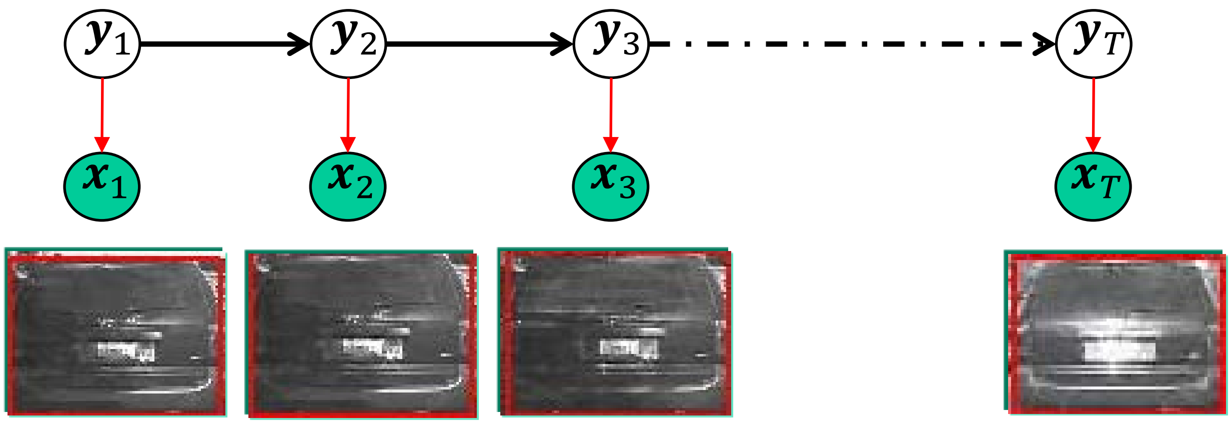


What we will learn?:

- *How to formulate probabilistically the problem.*
- *How to find both data moments* $\boldsymbol{E}[\boldsymbol{y}_i]$, $\boldsymbol{E}[\boldsymbol{y}\boldsymbol{y}^{\boldsymbol{T}}]$ *and parameters* θ

Summarize what we will study?

Sequential data (3 weeks):



What are the models?:

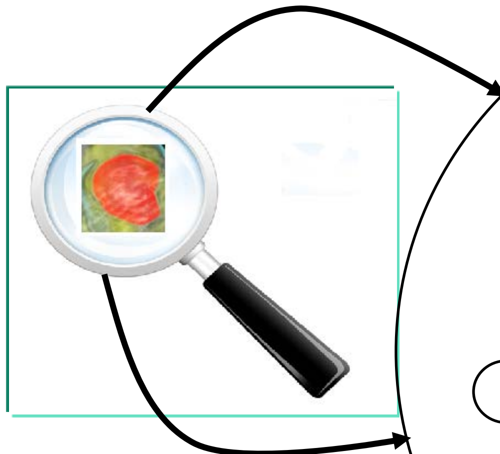
- *The Kalman filter (1 week)/ the particle filter (1 week).*
- *The Hidden Markov Model (1 week).*

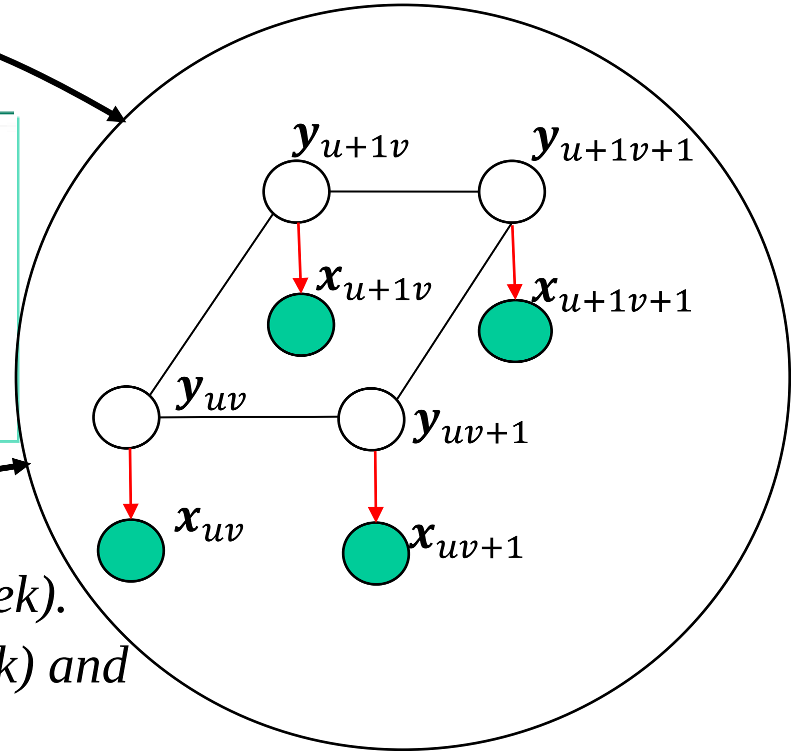
What we will learn?:

- *How to formulate probabilistically the problems and learn parameters.*

Summarize what we will study?

Spatial data (2-3 weeks):





What are the models?:

- *Gaussian Markov Random Fields (1 week).*
- *Discrete Markov Random Fields (1 week) and Mean Field Approximation.*

What we will learn?:

- *How to formulate probabilistically the problems and learn parameters.*

What are the tools we need?

- We need elements from differential/integral calculus using vectors and matrices
  (e.g., $\nabla_{\boldsymbol{W}} f(\boldsymbol{W}) = \dfrac{\partial f(\boldsymbol{W})}{d w_{ij}}$)

  ✓ *Matrix cookbook* (by Michael Syskind Pedersen & Kaare Brandt Petersen)

  ✓ Mike Brookes (EEE) has a nice page

  http://www.ee.ic.ac.uk/hp/staff/dmb/matrix/calculus.html

- We need linear algebra (matrix multiplications, matrix inversion etc). Special focus on eigenanalysis.

  ✓ Excellent book is *Matrix Computation by Gene H. Golub, & Charles F. Van Loan*

What are the tools we need?

- We need elements of optimization (mainly simple quadratic optimization problems with constraints which result to generalized eigenvalue problems). Refresh memory on how Lagrangian multipliers are used etc.

- We need tools from probability/statistics: random variable, probability density/mass function, marginalization
  - ✓ Assume pdf $p(x)$ the probability is computed $P(x \in A) = \int_A p(x)dx$
  - ✓ Marginal distributions $p(x) = \int_x p(x,y)dx$

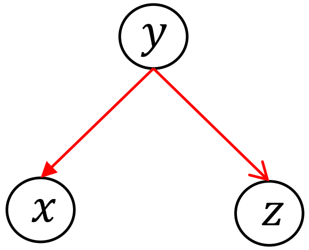
    $$p(y) = \int_y p(x,y)dy$$
  - ✓ First and second order moments $\mathrm{E}(\boldsymbol{x}) = \int_{\boldsymbol{x}} \boldsymbol{x}p(\boldsymbol{x})d\boldsymbol{x}$

    $$\mathrm{E}(\boldsymbol{x}\boldsymbol{x}^T) = \int_{\boldsymbol{x}} \boldsymbol{x}\boldsymbol{x}^T p(\boldsymbol{x})d\boldsymbol{x}$$

What are the tools we need?

- Bayes rule and conditional independence.

$$p(x, y) = p(x|y)p(y)$$ Bayes rule



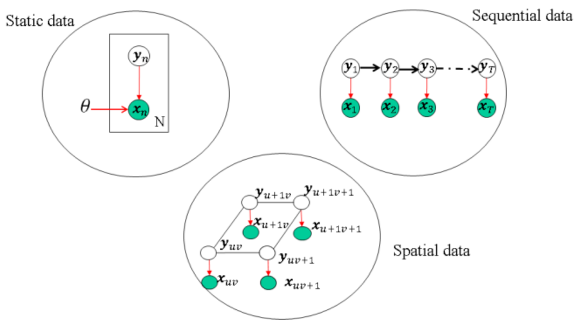
Conditional independence

$$p(x, z, y) = p(x|y)p(z|y)p(y)$$

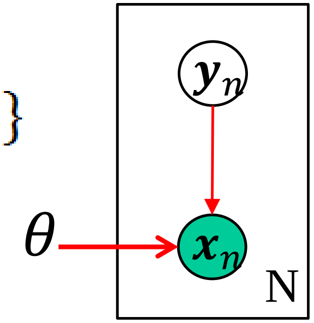
- Finally, we need tools from algorithms recursion and dynamic programming

What to have always in mind?

- What is my model?



- What are my model's parameters? $\theta = \{W, \mu, \sigma^2\}$



- How do I find them? Maximum Likelihood
  Expectation Maximization

- How do I use the model?

Assignments

- Assessment (90% from written exams & 10% from assignments)

- Two assignments:

  One will be given next week and should be delivered by 21st of February

  The second will be given 24th of February and should be delivered 17th of March

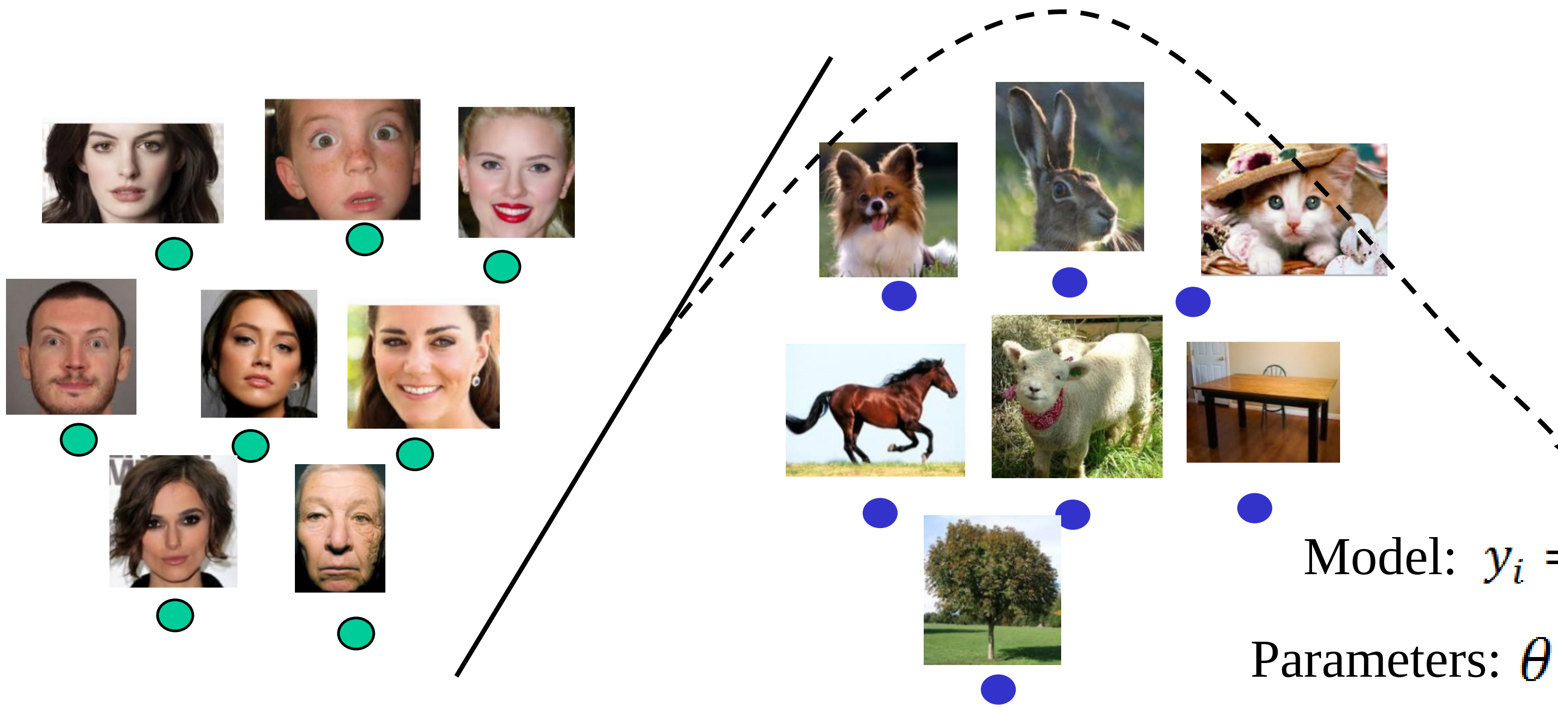
Adv. Statistical Machine Learning – Lectures

- Lecture 1-2: Introduction
- Lecture 3-4: A primer on calculus, linear algebra, probability/statistics
- Lecture 5-6: Deterministic Component Analysis (1)
- Lecture 7-8: Deterministic Component Analysis (2)
- Lecture 9-10: Deterministic Component Analysis (3)
- Lecture 11-12: Probabilistic Principal Component Analysis
- Lecture 13-14: Sequential Data: Kalman Filter (1)
- Lecture 15-16: Sequential Data: Kalman Filter (2)
- Lecture 17-18: Sequential Data: Hidden Markov Model (1)
- Lecture 18-19: Sequential Data: Hidden Markov Model (2)
- Lecture 19-20: Sequential Data: Particle Filtering (1)
- Lecture 20-21: Sequential Data: Particle Filtering (2)
- Lecture 22-23: Spatial Data: Gaussian Markov Random Field (GMRF)
- Lecture 23-24: Spatial Data: GMRF (2)
- Lecture 25-28: Spatial Data: Discrete MRF (Mean Field)

Machine Learning (models/parameters)

What does the machine learn in each application (parameters of a model)?

- Object detection: learn classifier, i.e. a function (design of classifiers is covered in Neural Computation)



Model: $y_i = \boldsymbol{w}^T \boldsymbol{x}_i + \boldsymbol{b}$

Parameters: $\theta = \{\mathbf{w}, b\}$