Deep Learning

What is it? And what are we doing about it?

Neil D. Lawrence
Sheffield Institute of Translational Neuroscience and
Department of Computer Science, University of Sheffield, U.K.

Departmental Seminar: DCS, University of Sheffield

20th March 2013

Scientists See Promise in Deep-Learning Programs

Hao Zhang/The New York Times

A voice recognition program translated a speech given by Richard F. Rashid, Microsoft's top scientist, into Mandarin Chinese.

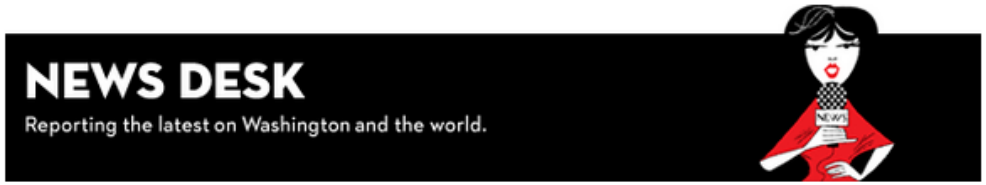
By JOHN MARKOFF
Published: November 23, 2012

Using an artificial intelligence technique inspired by theories about how the brain recognizes patterns, technology companies are reporting startling gains in fields as diverse as computer vision, speech recognition and the identification of promising new molecules for designing drugs.

FACEBOOK
TWITTER
GOOGLE+
SAVE
E-MAIL
SHARE
PRINT

点击查看本文中文版。

The advances have led to widespread enthusiasm among researchers who design software to perform human

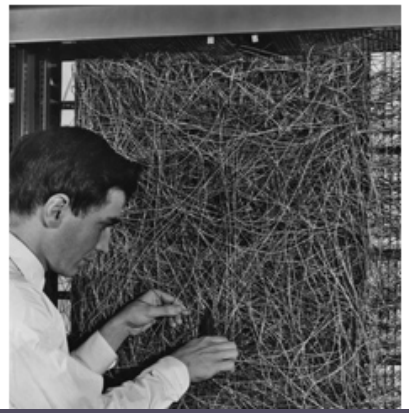


« How Susan Rice Sees the World | Main | Moral Machines »

NOVEMBER 25, 2012

IS "DEEP LEARNING" A REVOLUTION IN ARTIFICIAL INTELLIGENCE?

POSTED BY GARY MARCUS



Can a new technique known as deep learning revolutionize artificial intelligence, as yesterday's front-page article at the New York *Times* suggests? There is good reason to be excited about deep learning, a sophisticated "machine learning" algorithm that far exceeds many of its predecessors in its abilities to recognize syllables and images. But there's also good reason to be skeptical. While the *Times* reports that "advances in an artificial intelligence technology that can recognize patterns

Google To Expand Knowledge Graph Through Hire Of Geoffrey Hinton

PREV STORY | NEXT STORY

Mar 14, 2013 • 8:23 am | (10)
by Barry Schwartz | Filed Under Google Search Engine

If I had to place one search priority above all else, I'd say right now, Google's most ambitious project is the knowledge graph. Yea, they are pushing Google+ big time, but the knowledge graph is a level above all of that technically.

Of course, Google has an outstanding team working on this project lead by one of the smartest people I've ever met Amit Singhal.

To take the knowledge graph to the next level, Google has hired/acquired Geoffrey Hinton and his team at DNNresearch. Geoffrey posted a note on his Google+ page about it:

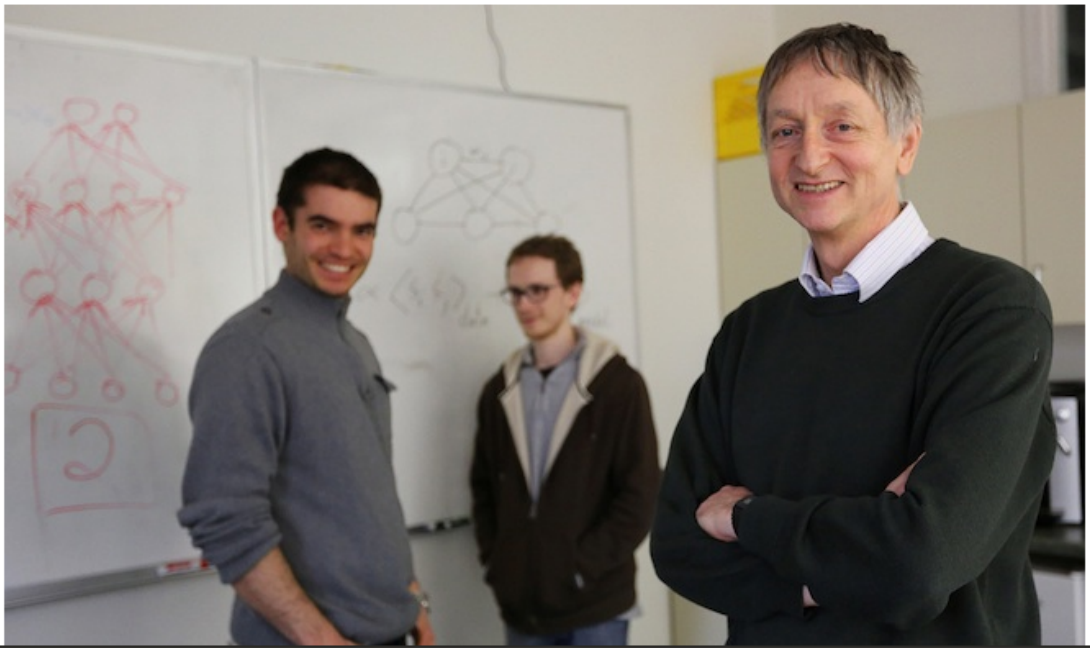
> Last summer, I spent several months working with Google's Knowledge team in Mountain View, working with Jeff Dean and an incredible group of scientists and engineers who have a real shot at making spectacular progress in machine learning. Together with two of my recent graduate students, Ilya Sutskever and Alex Krizhevsky (who won the 2012 ImageNet competition), I am betting on Google's team to be the epicenter of future breakthroughs. That means we'll soon be joining Google to work with some of the smartest engineering minds to tackle some of the biggest challenges in computer science. I'll remain part-time at the University of Toronto, where I still have a lot of excellent graduate students, but at Google I will get to see what we can do with very large-scale computation.

I know we just scratched the surface of the knowledge graph and I am excited to see where it takes us in the future.

I am just glad I don't have to figure out how to get us there. I get to just sit and enjoy the ride.

Google Hires Brains that Helped Supercharge Machine Learning

BY ROBERT MCMILLAN 03.13.13 6:30 AM



MOST RECENT WIRED POSTS


Jawbone's Up Fitness Band Is Now Android-Compatible


Review: Ecovacs Winbot, a Window-Cleaning Robot


Sherlock, Professor X and Margaery Tyrell Team for Neil Gaiman Radio Play


Video: Robo-Chopper Dives and Grabs Objects Like a Bird of Prey

Google Hires Geoffrey Hi | Google Hires Brains that | Geoffrey Hinton - Google+

https://plus.google.com/102889418997957626067/posts

My Boosters | LastPass - Dow... | Add to TripIt | Proverbi napol... | IEEE Xplore - On... | Google Maps | Other Bookmarks

+Neil Search Images Maps Play YouTube News Gmail Drive Calendar More

Google+ Neil Lawrence 0 + Share

Home
Profile
Explore
Events
Photos
Communities
NEW Find people
Local
Games
Hangouts
More

Geoffrey Hinton

1,734 have him in circles

ML People

23 in common

About **Posts** Photos Videos

Geoffrey Hinton 12 Mar 2013 - Public

Last summer, I spent several months working with Google's Knowledge team in Mountain View, working with Jeff Dean and an incredible group of scientists and engineers who have a real shot at making spectacular progress in machine learning. Together with two of my recent graduate students, Ilya Sutskever and Alex Krizhevsky (who won the 2012 ImageNet competition), I am betting on Google's team to be the epicenter of future breakthroughs. That means we'll soon be joining Google to work with some of the smartest engineering minds to tackle some of the biggest challenges in computer science. I'll remain part-time at the University of Toronto, where I still have a lot of excellent graduate students, but at Google I will get to see what we can do with very large-scale computation.

+1 +418 167

64 comments

Reza Samahin 15 Mar 2013

+Geoffrey Hinton congrats to you and your team from an old UofT eng grad. Wish I were young again to contribute to your endeavour.

Add a comment...

43 IN HIS CIRCLES

 George Dahl Add

 David Reichert Add

 Nitish Srivastava Add

 Jacqueline Ford Add

 Aaron Hertzmann Add

 Navdeep Jaitly Add

23 IN COMMON WITH YOU

1,734 HAVE

Chat

direction for further research.

11.1. HAVE WE THROWN THE BABY OUT WITH THE BATH WATER?

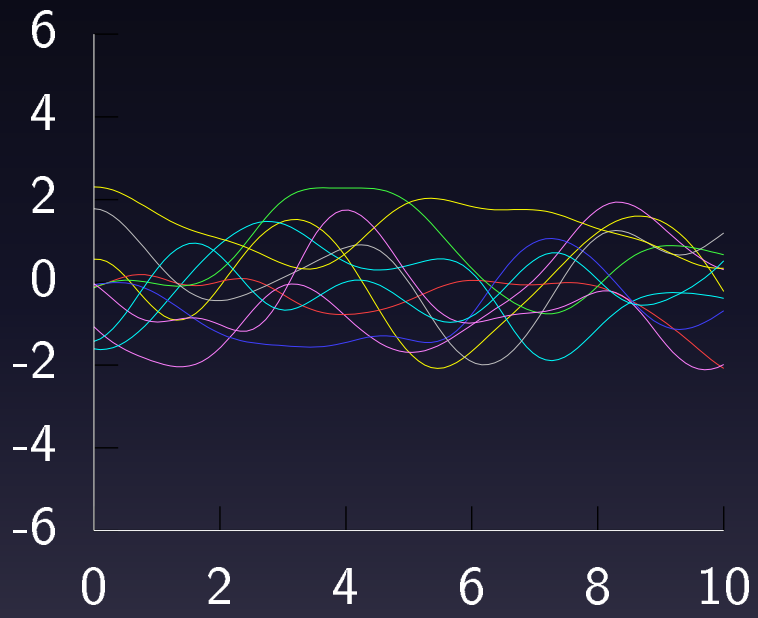
According to the hype of 1987, neural networks were meant to be intelligent models which discovered features and patterns in data. Gaussian processes in contrast are simply smoothing devices. How can Gaussian processes possibly replace neural networks? What is going on?

I think what the work of Williams and Rasmussen (1996) shows is that many real–world data modelling problems are perfectly well solved by sensible smoothing methods. The most interesting problems, the task of feature discovery for example, are not ones which Gaussian processes will solve. But maybe multilayer perceptrons can't solve them either. On the other hand, it may be that the limit of an infinite number of hidden units, to which Gaussian processes correspond, was a bad limit to take; maybe we should backtrack, or modify the prior on neural network parameters, so as to create new models more interesting than Gaussian processes. Evidence that this infinite limit has lost something compared with finite neural networks comes from the observation that in a finite neural network with more than one output, there are non–trivial correlations between the outputs (since they share inputs from common hidden units); but in the limit of an infinite number of hidden units, these correlations vanish. Radford Neal has suggested the use of non–Gaussian priors in networks with multiple hidden layers. Or perhaps a completely fresh start is needed, approaching the problem of machine learning from a paradigm different from the supervised feedforward mapping.

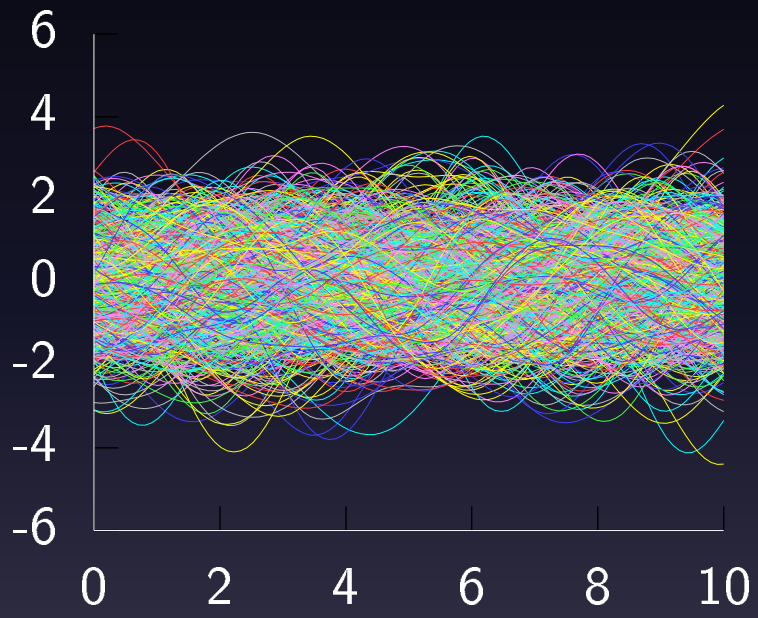
Structure of Priors

MacKay: NIPS Tutorial 1997 “Have we thrown out the baby with the bathwater?” (Published as MacKay, 1998) Also noted by (Wilson et al., 2012)

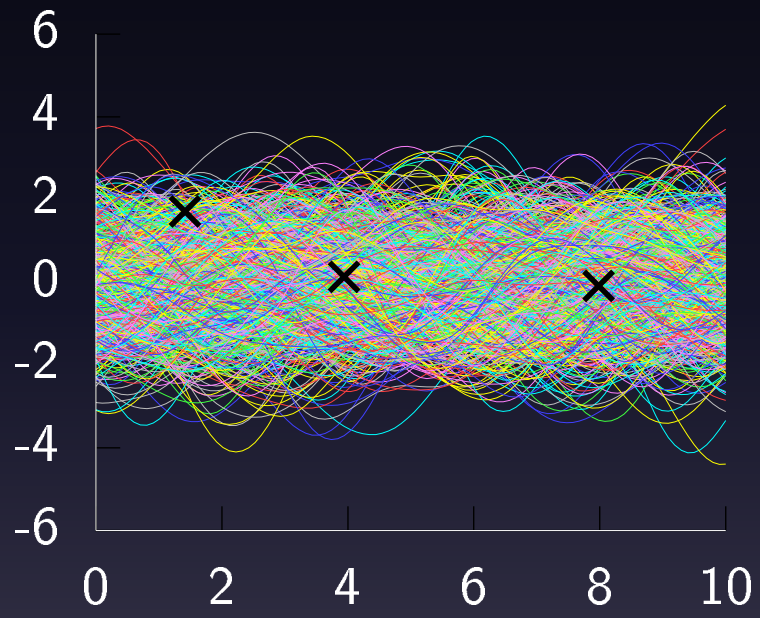
Gaussian Processes



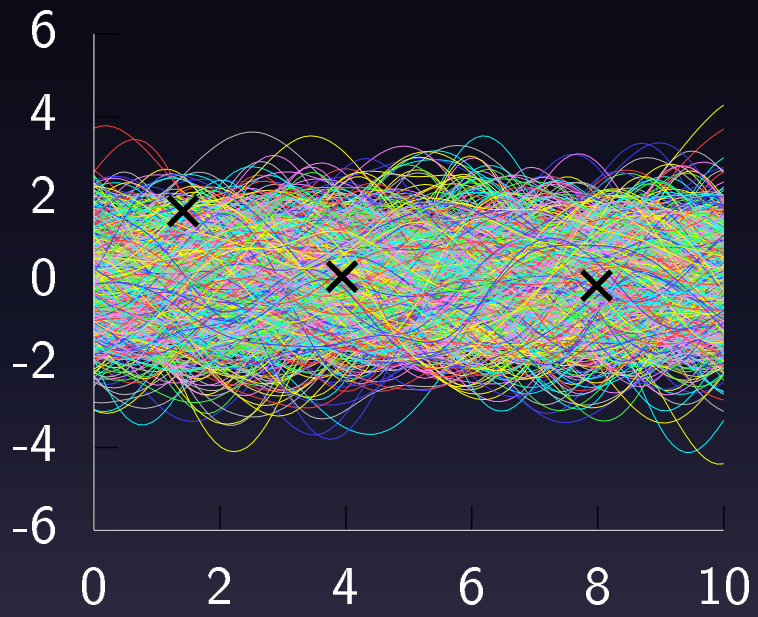
Gaussian Processes

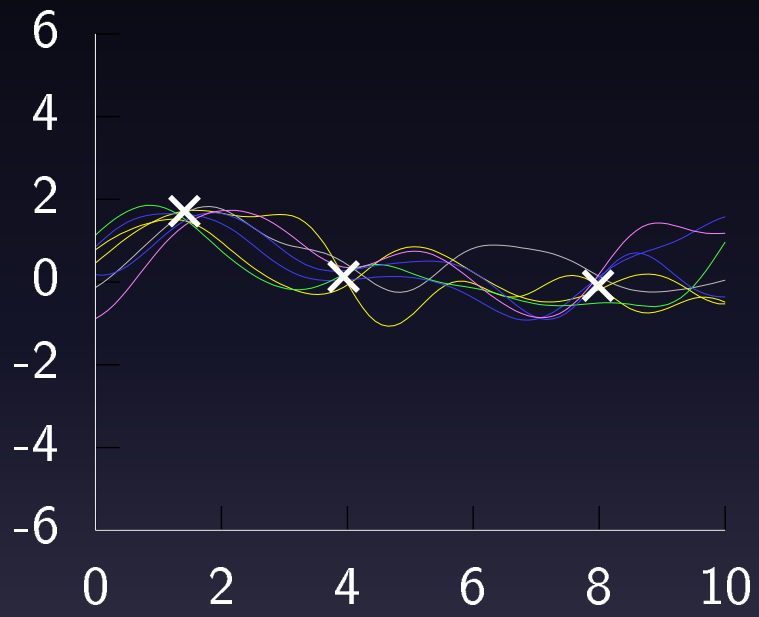


Gaussian Processes



Gaussian Processes





Motivation for Deep Learning

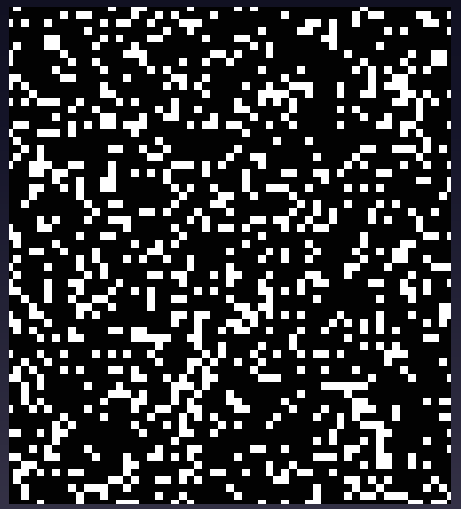
USPS Data Set Handwritten Digit

- 3648 Dimensions
  - 64 rows by 57 columns
  - Space contains more than just this digit.
  - Even if we sample every nanosecond from now until the end of the universe, you won't see the original six!

Motivation for Deep Learning

USPS Data Set Handwritten Digit

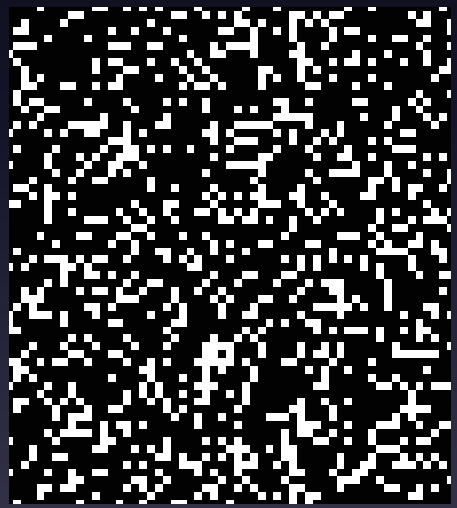
- 3648 Dimensions
  - 64 rows by 57 columns
  - Space contains more than just this digit.
  - Even if we sample every nanosecond from now until the end of the universe, you won't see the original six!



Motivation for Deep Learning

USPS Data Set Handwritten Digit

- 3648 Dimensions
  - 64 rows by 57 columns
  - Space contains more than just this digit.
  - Even if we sample every nanosecond from now until the end of the universe, you won't see the original six!



Motivation for Deep Learning

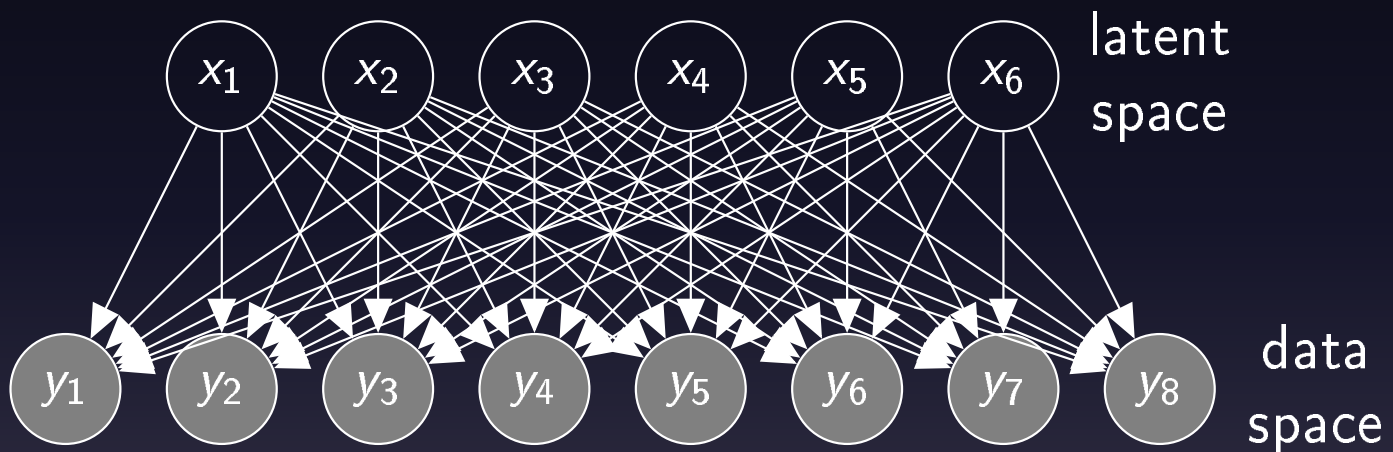
USPS Data Set Handwritten Digit

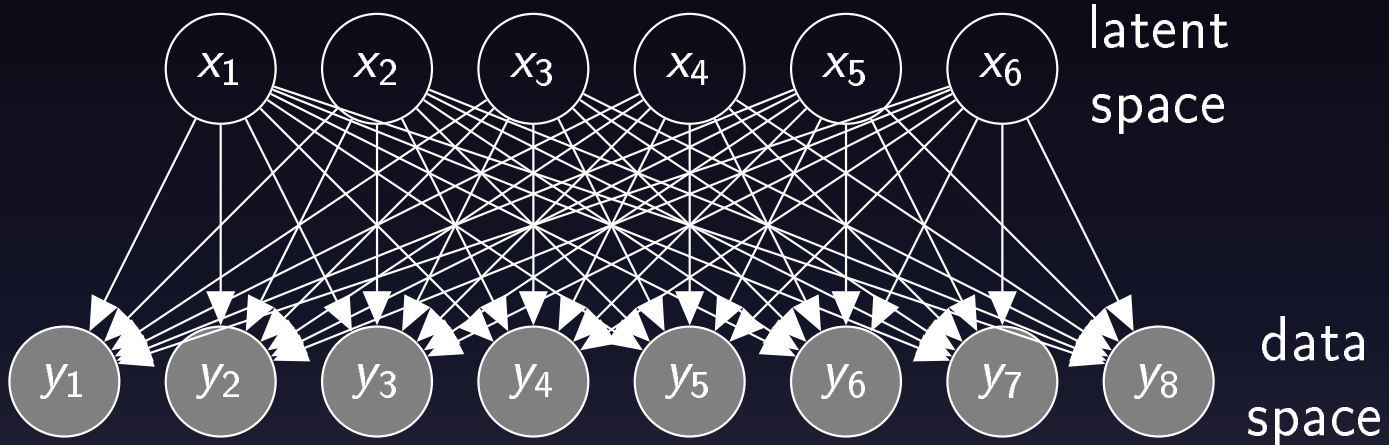
- 3648 Dimensions
  - 64 rows by 57 columns
  - Space contains more than just this digit.
  - Even if we sample every nanosecond from now until the end of the universe, you won't see the original six!

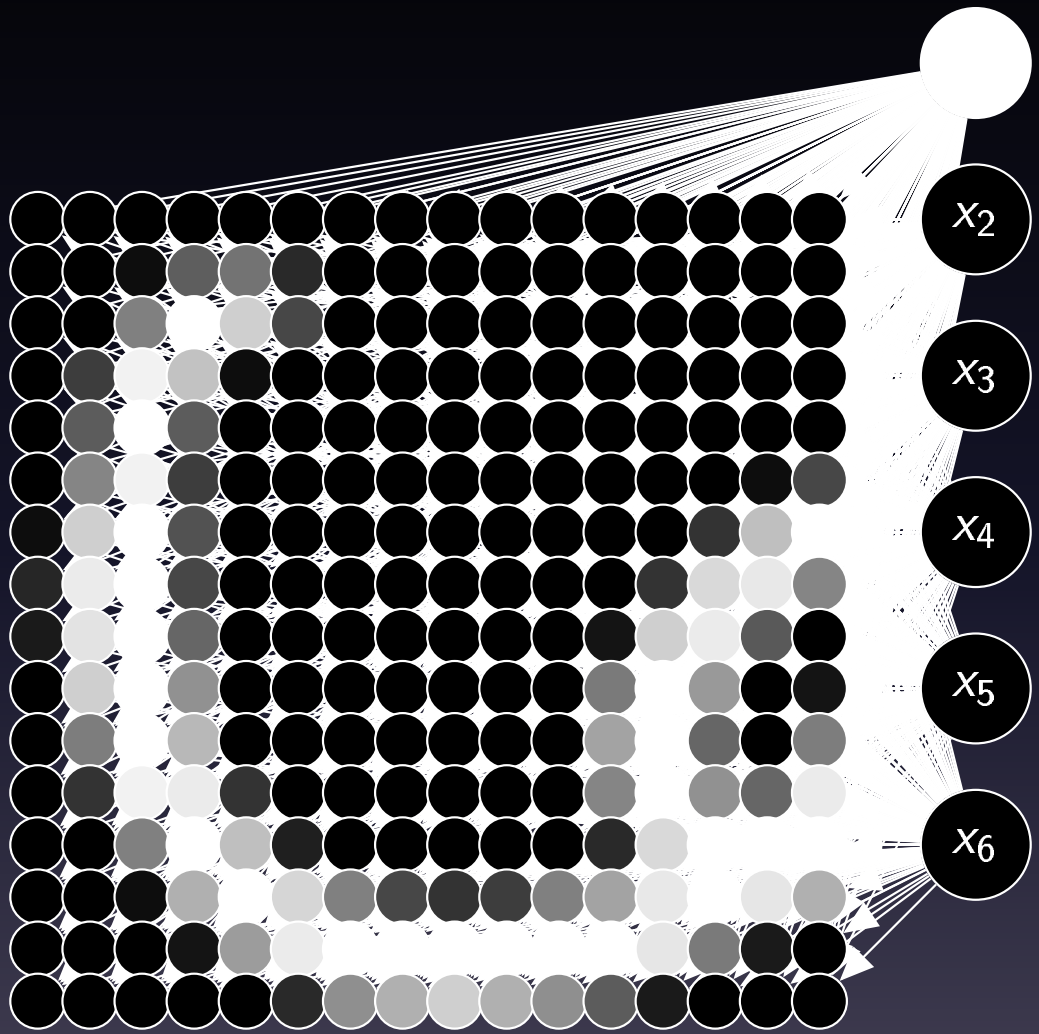
Template Model of Digits

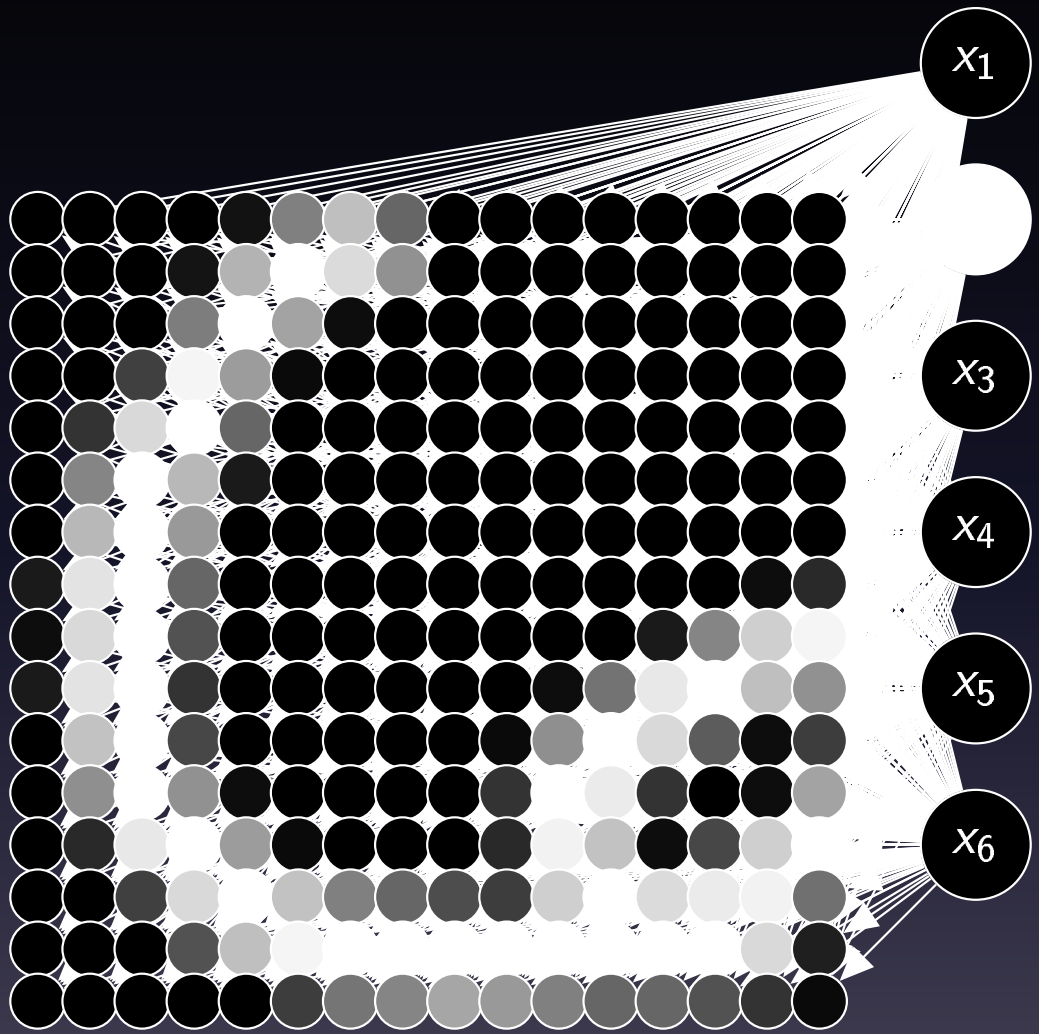
- Design a set of 'latent' features, which generate the 6.
- Global template: memorize data set.

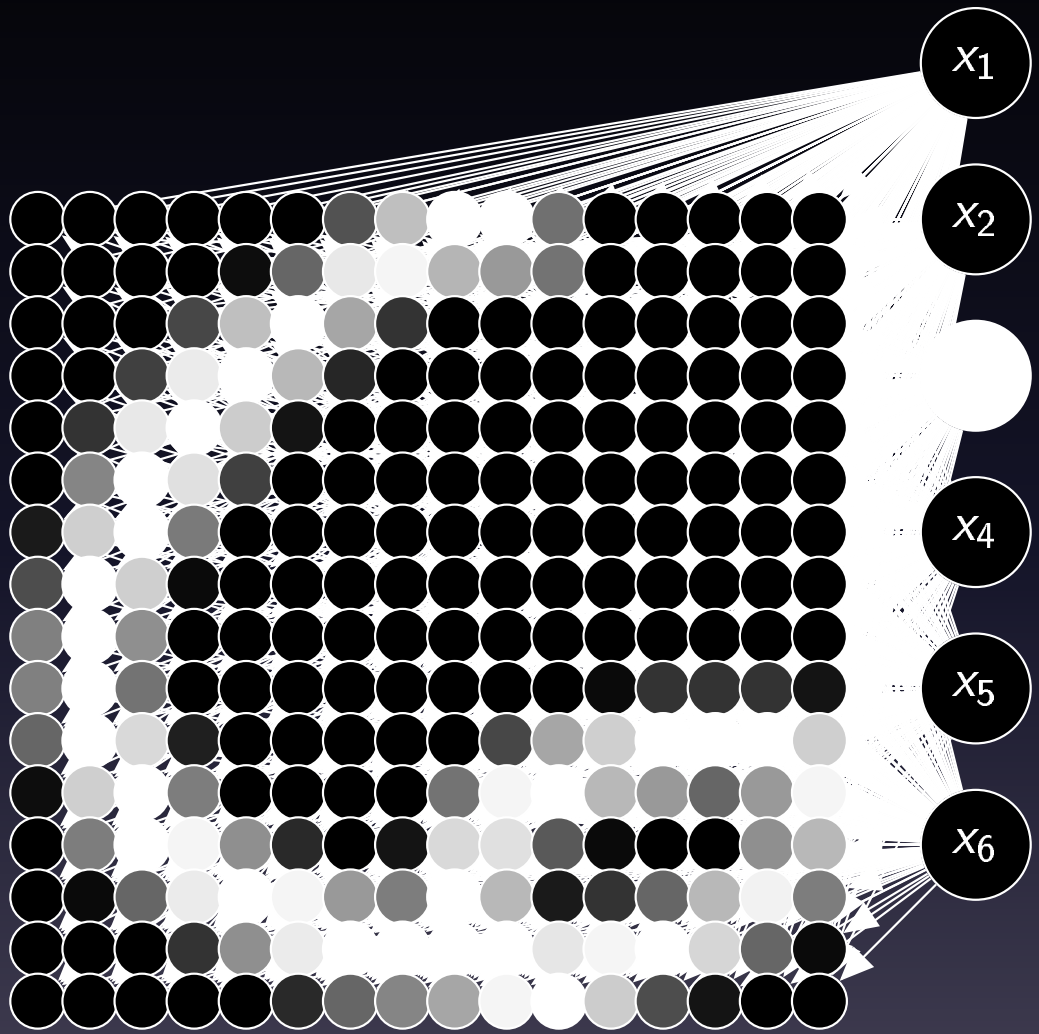
Latent Variable Model

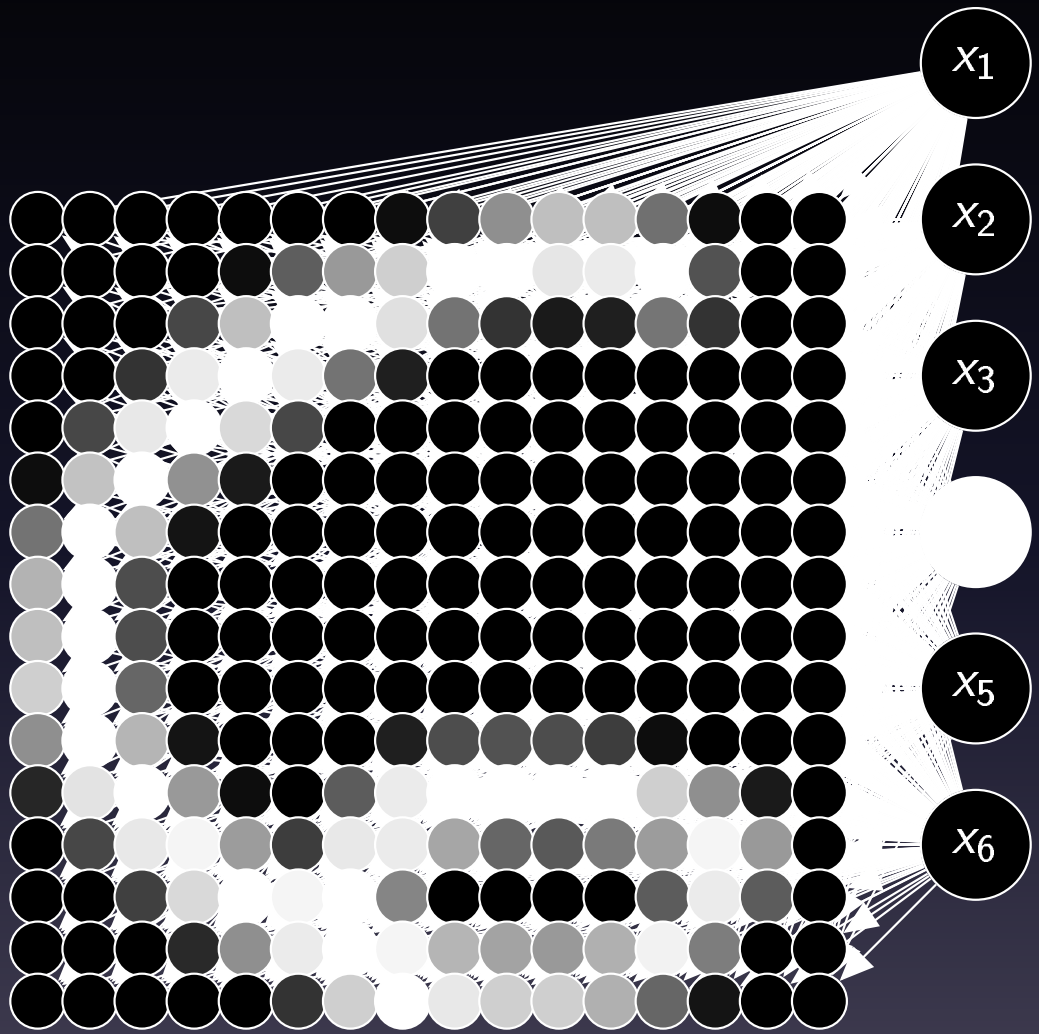


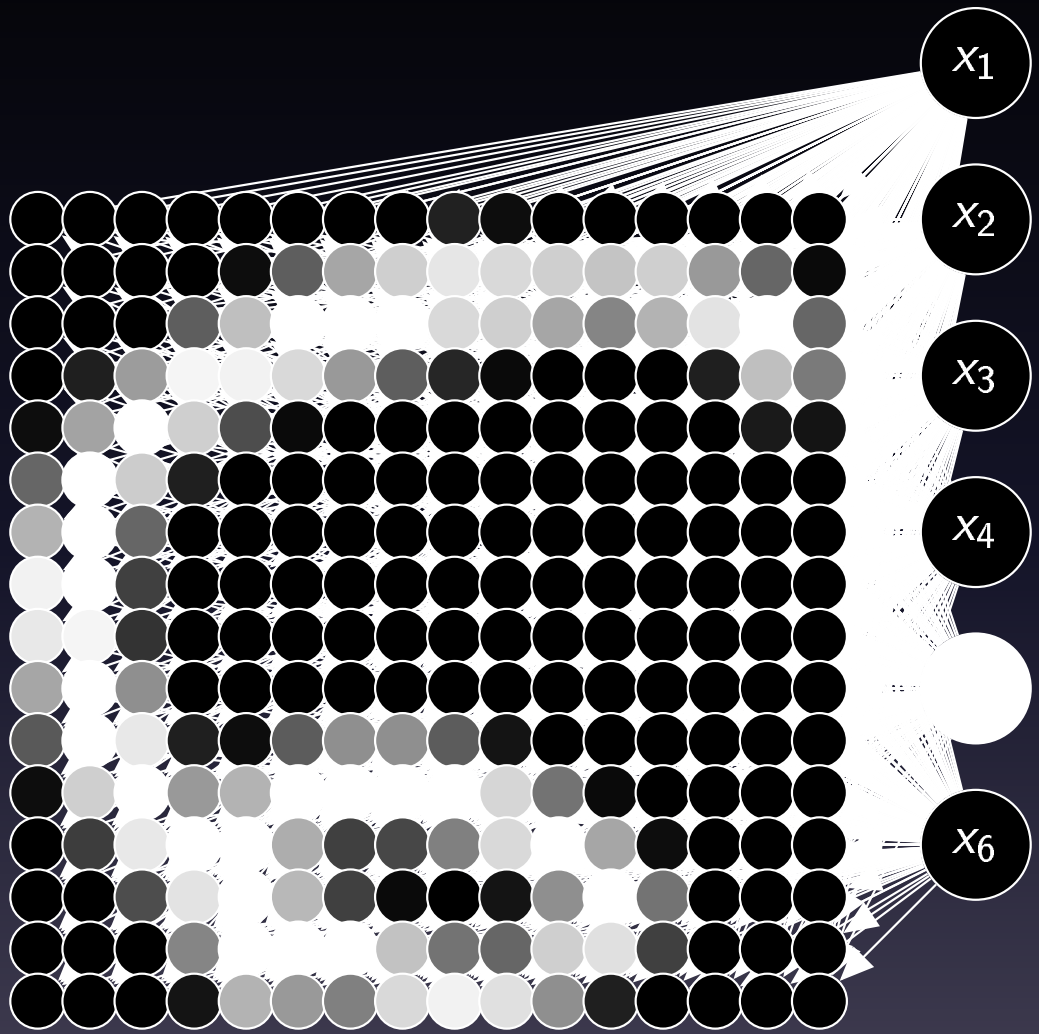

x_1
x_2
x_3
x_4
x_5
x_6
latent space
y_1
y_2
y_3
y_4
y_5
y_6
y_7
y_8
data space

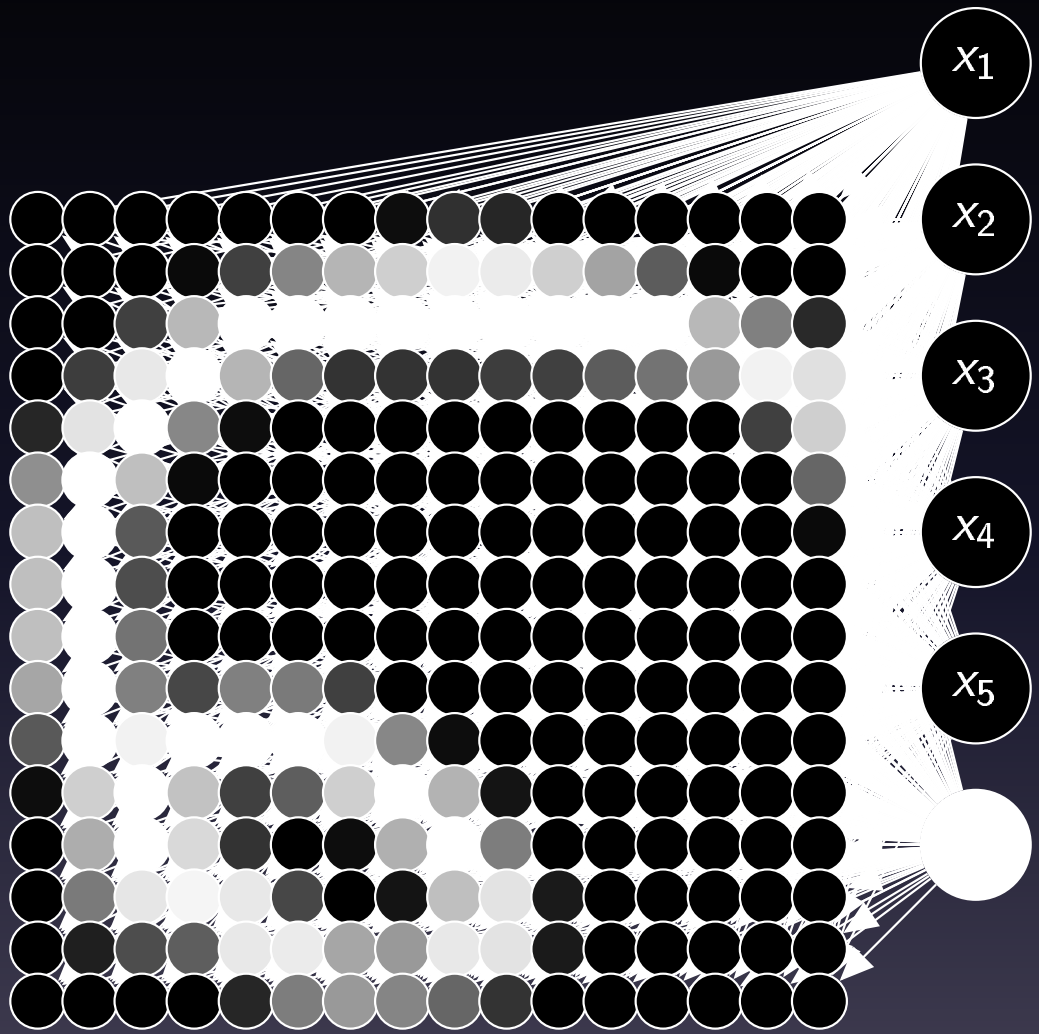

x_2
x_3
x_4
x_5
x_6


x_1
x_3
x_4
x_5
x_6


x_1
x_2
x_4
x_5
x_6


x_1
x_2
x_3
x_5
x_6


x_1
x_2
x_3
x_4
x_6


x_1
x_2
x_3
x_4
x_5

Template Matching

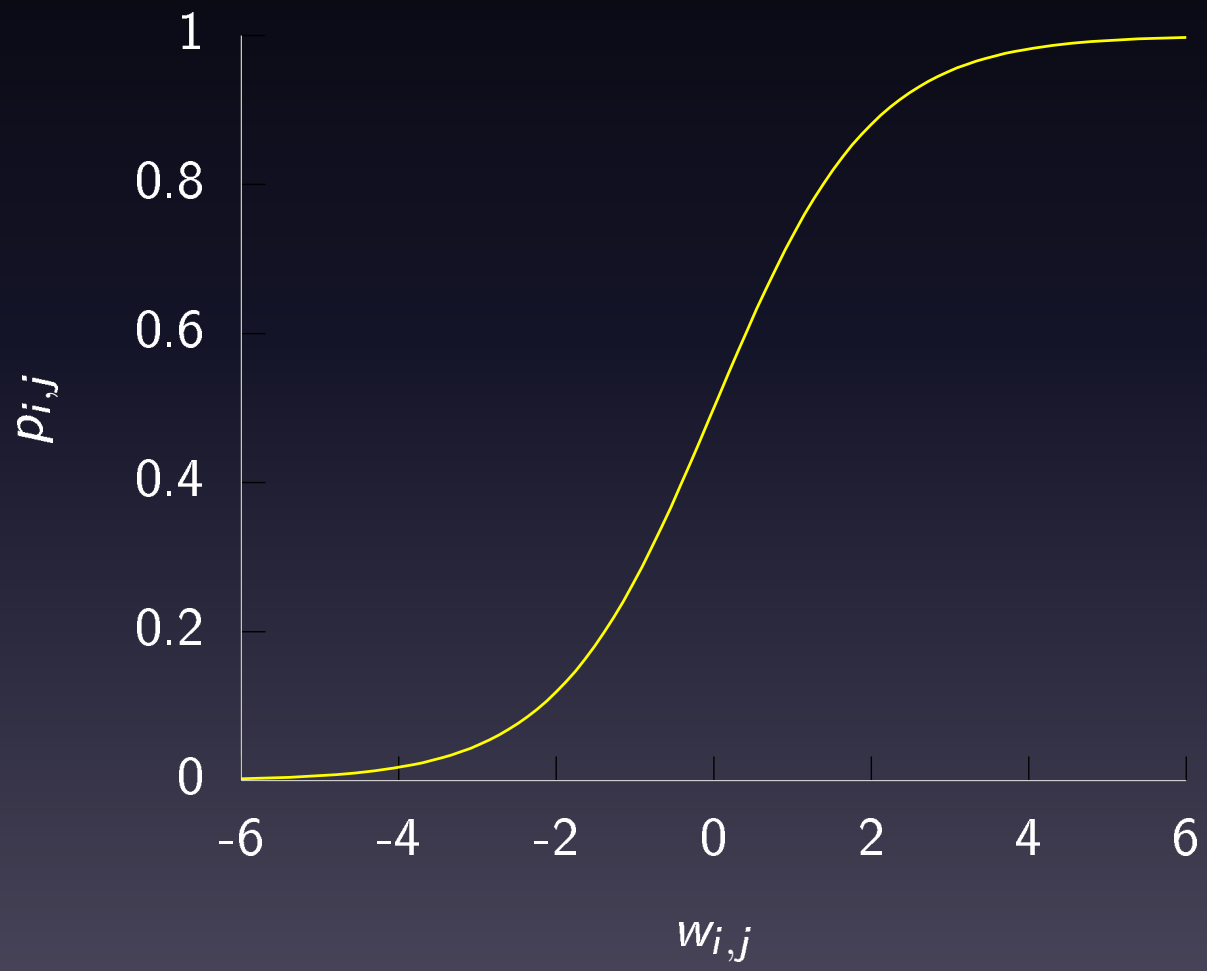
- Each latent node associated with a 'template' digit.
- If as many nodes as data then model is like 'nearest neighbour' with a particular distance measure.
- If less nodes than data then model is like a mixture of Bernoulli distributions.
- What if we allow several nodes to be switched on together?

Templates to Features

- In template matching ith node had an associated set of probabilities, $\mathbf{p}_i$.
- These probabilities can be reshaped into a matrix and sampled from to see the sixes.
- If the ith node is on the ith vector of probabilities is used.
- What if the ith node and the kth node are on?
    - How do we combine $\mathbf{p}_i$ and $\mathbf{p}_k$ to give probabilities of pixels?

Squashing Function

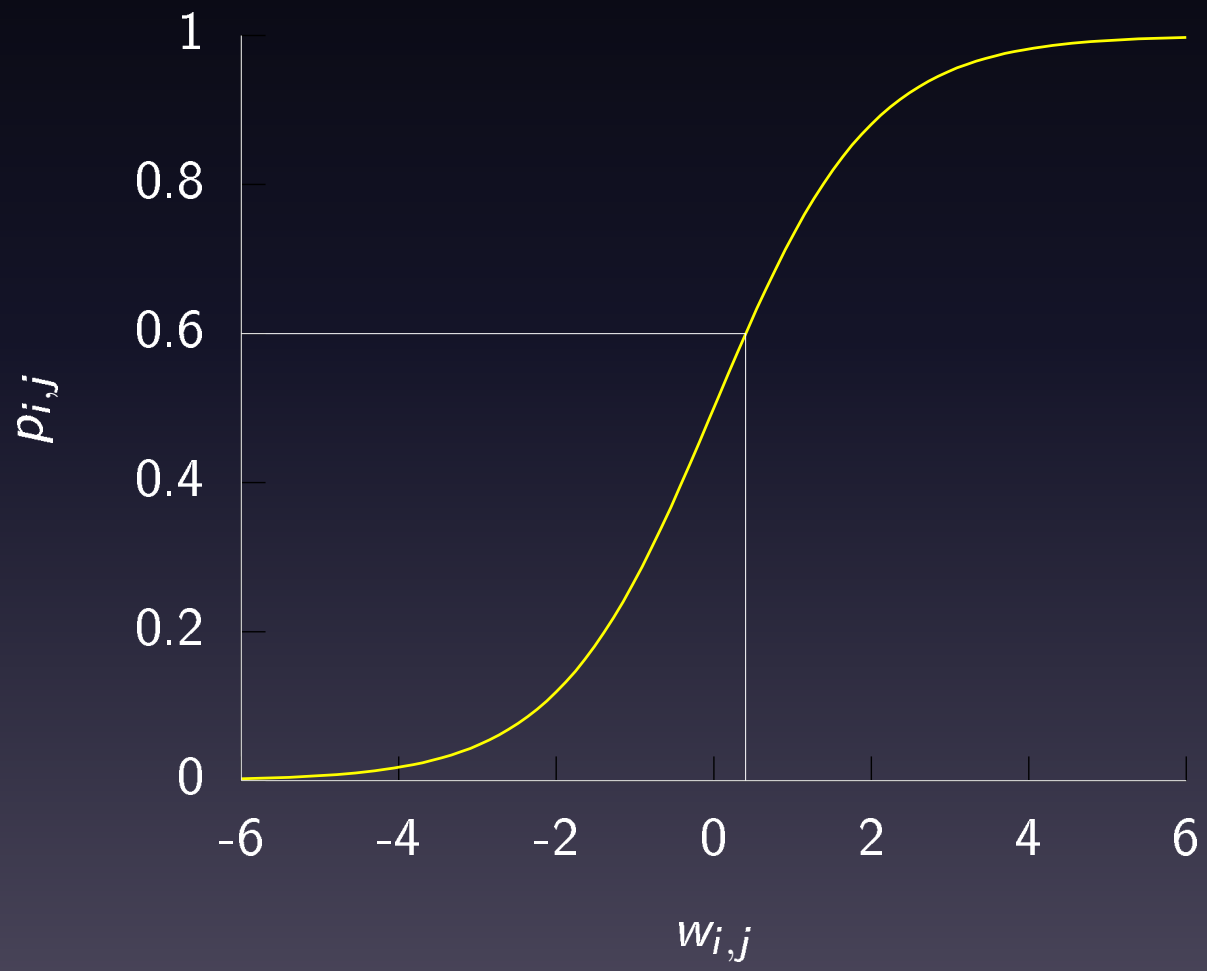
- One solution is to first reparameterise $p_{i,j}$ as a squashing function,



- For example the sigmoid function.

Squashing Function

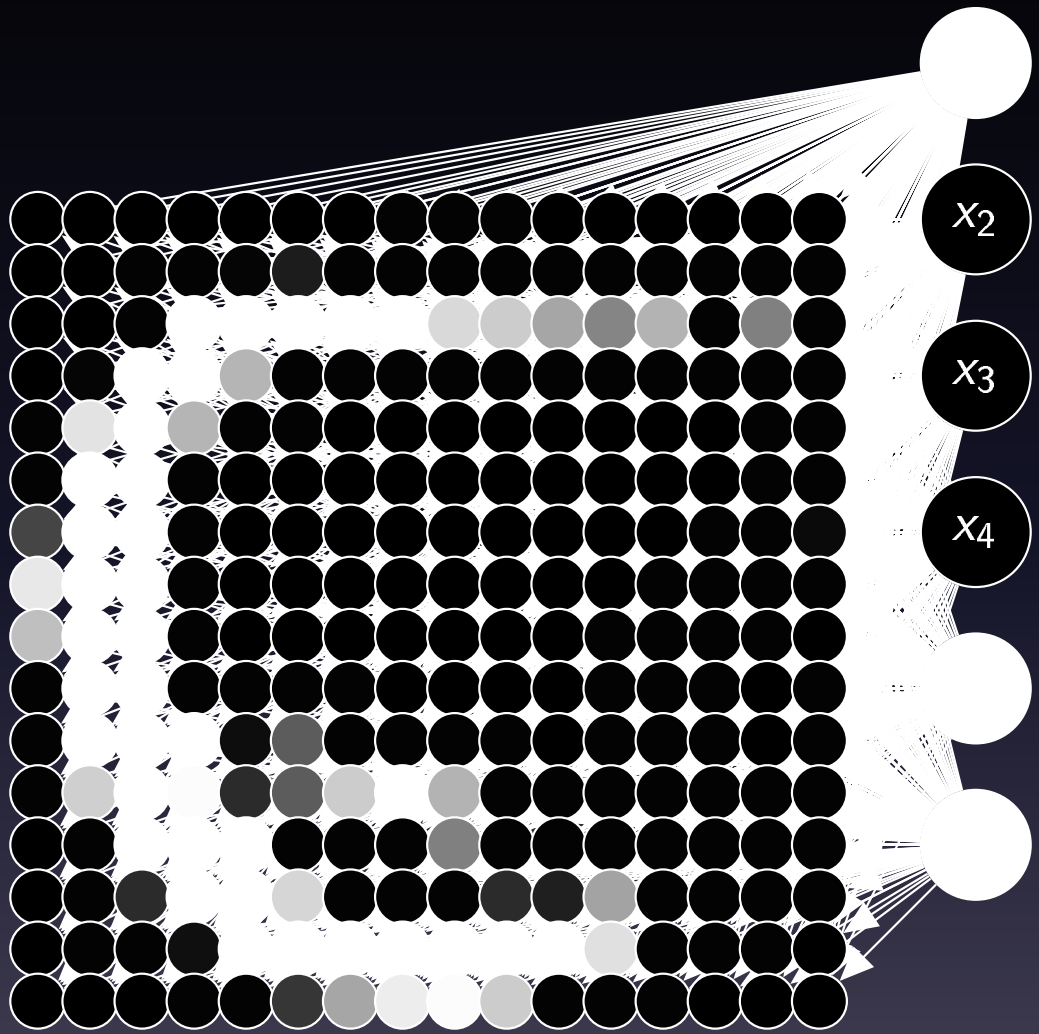
- One solution is to first reparameterise $p_{i,j}$ as a squashing function,

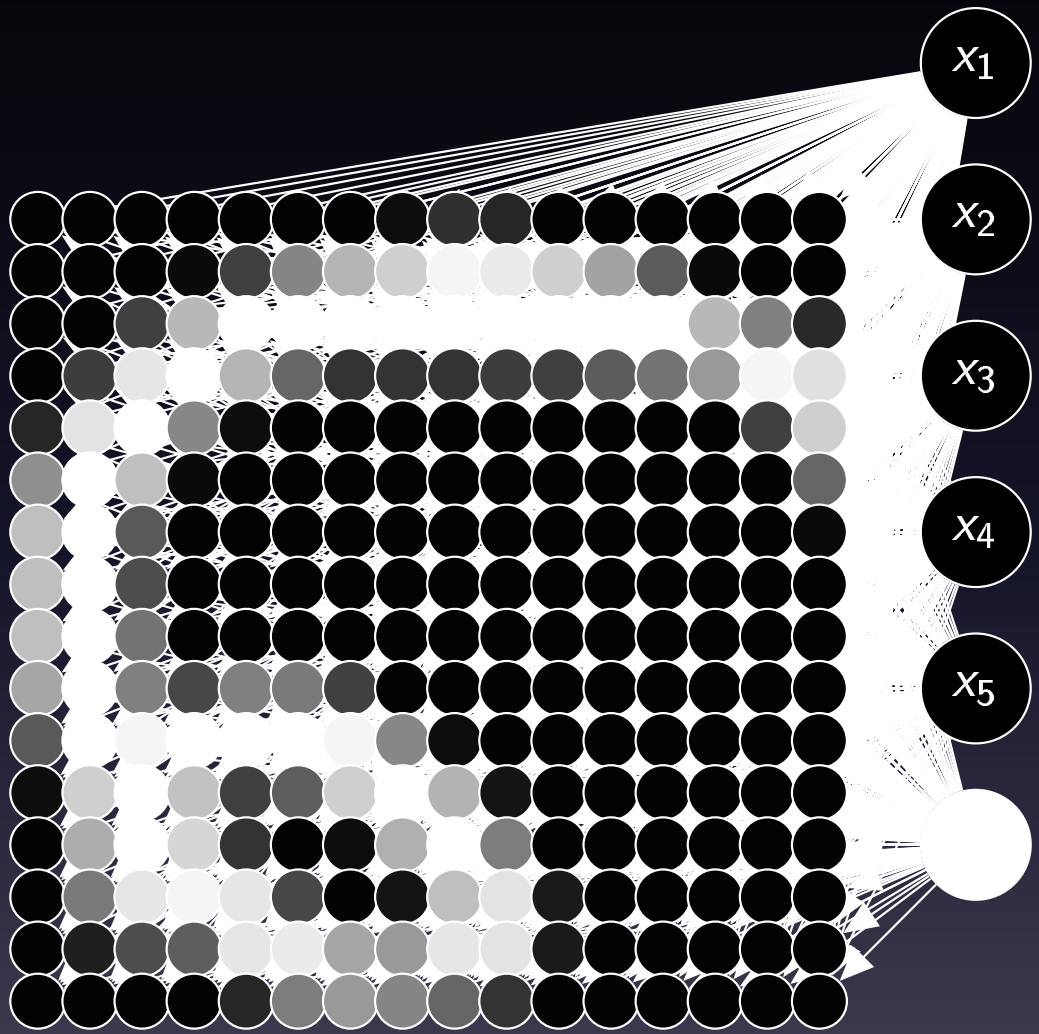


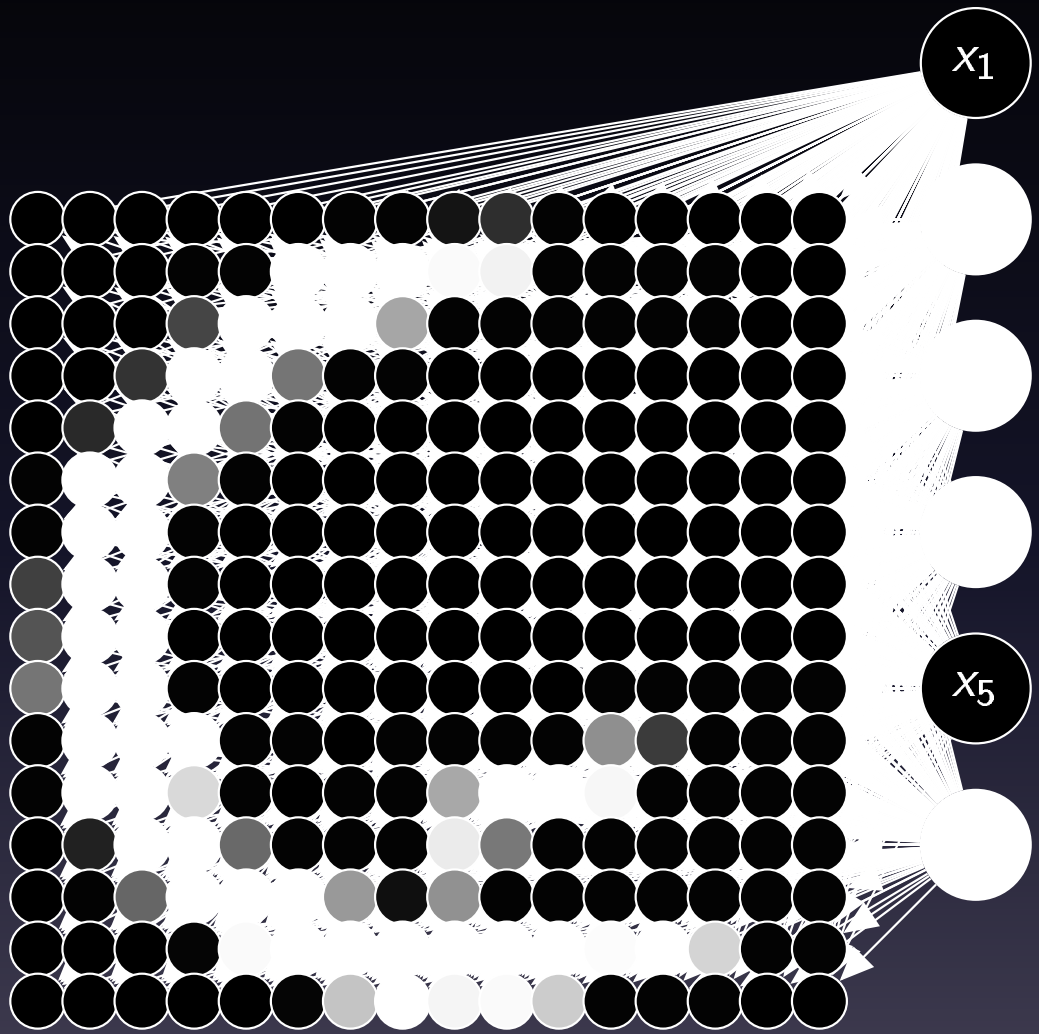
- For example the sigmoid function.

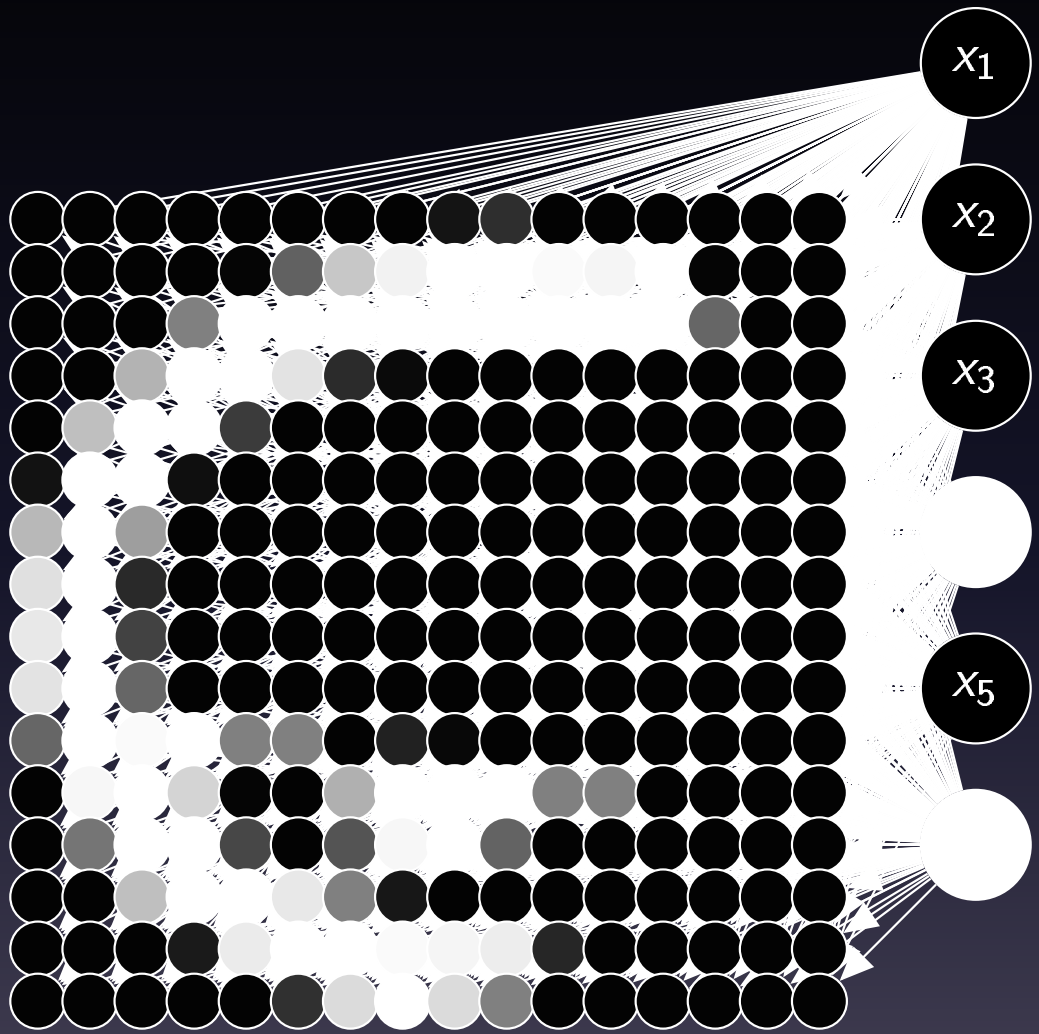
Addition Before Squashing

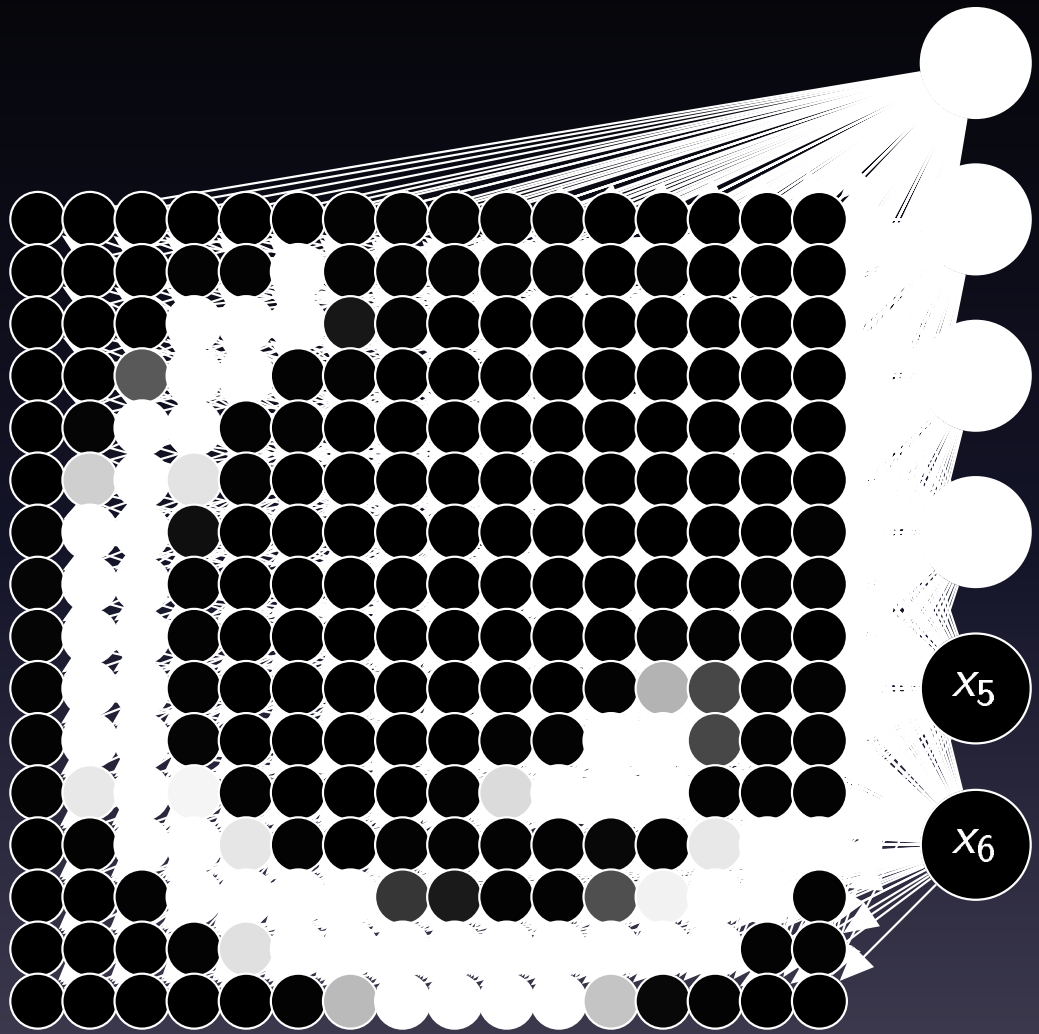
- Example: if latent node 1 and 6 are on.
- Can't add $p_{:,1}$ to $p_{:,6}$ to obtain probability that node is on.
- Instead add $w_{:,1}$ to $w_{:,6}$ and push through squashing function.
  - In general for $\mathbf{p}_{i,:}$ compute $\mathbf{W}\mathbf{x}_{i,:}$.
  - Then $p_{i,j} = \sigma(\mathbf{w}_{j,:}^\top \mathbf{x}_{i,j})$ where $\sigma(\cdot)$ is the sigmoid function.

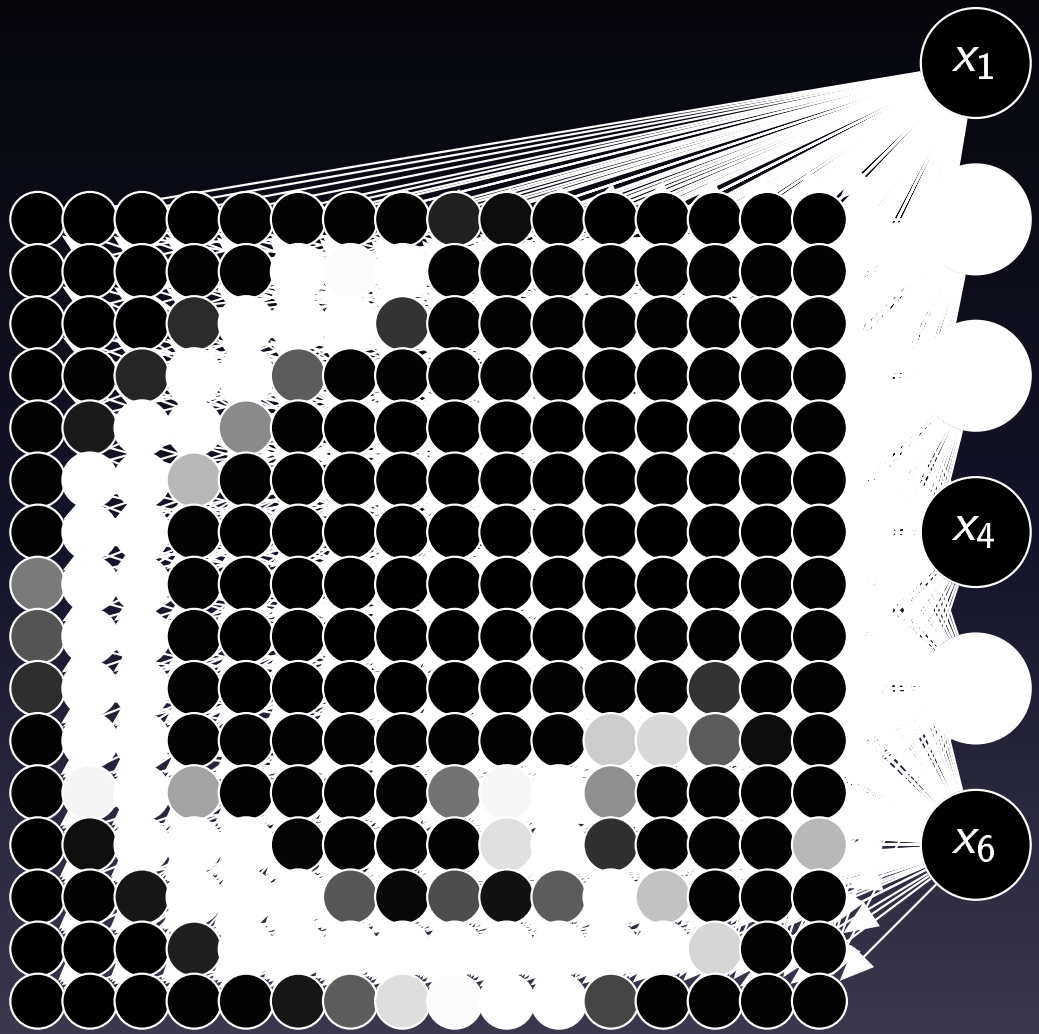

x_2
x_3
x_4


x_1
x_2
x_3
x_4
x_5


x_1
x_5

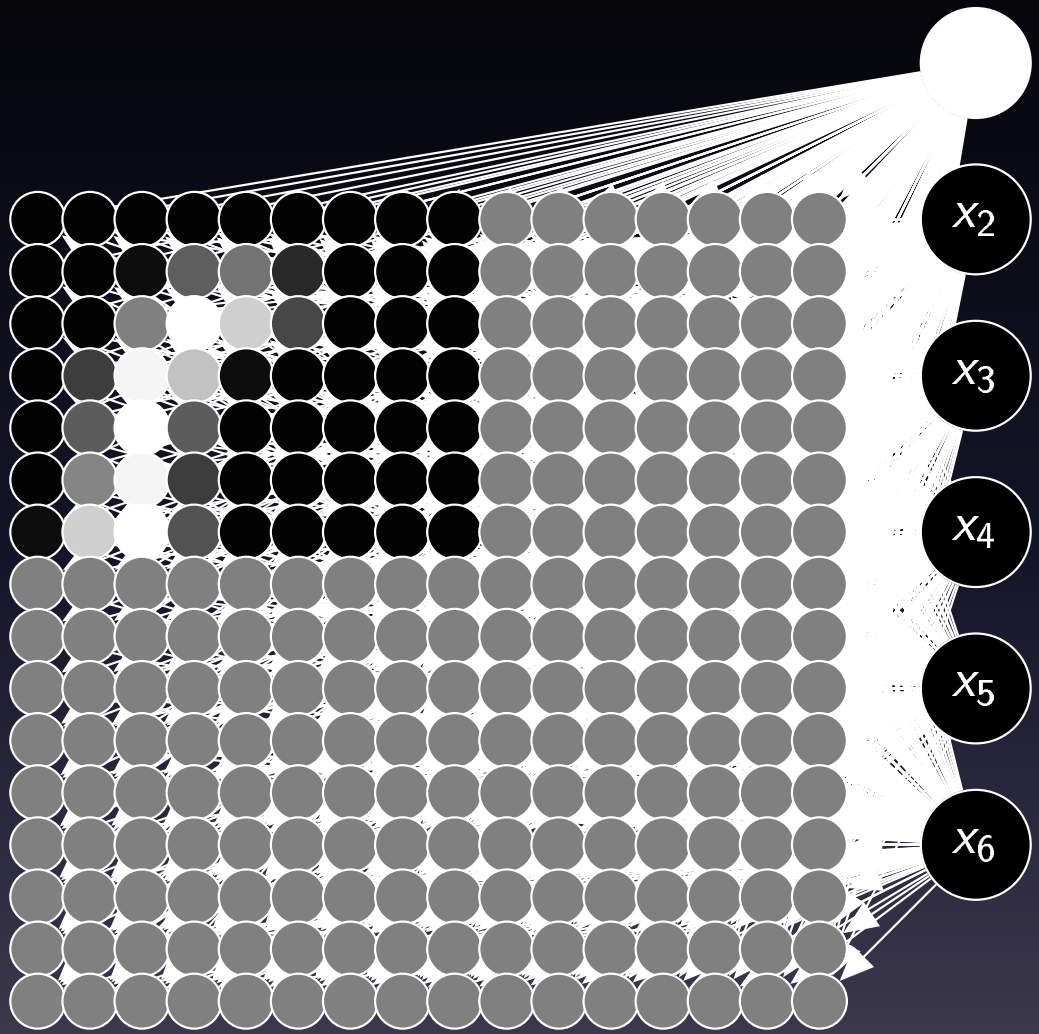

x_1
x_2
x_3
x_5

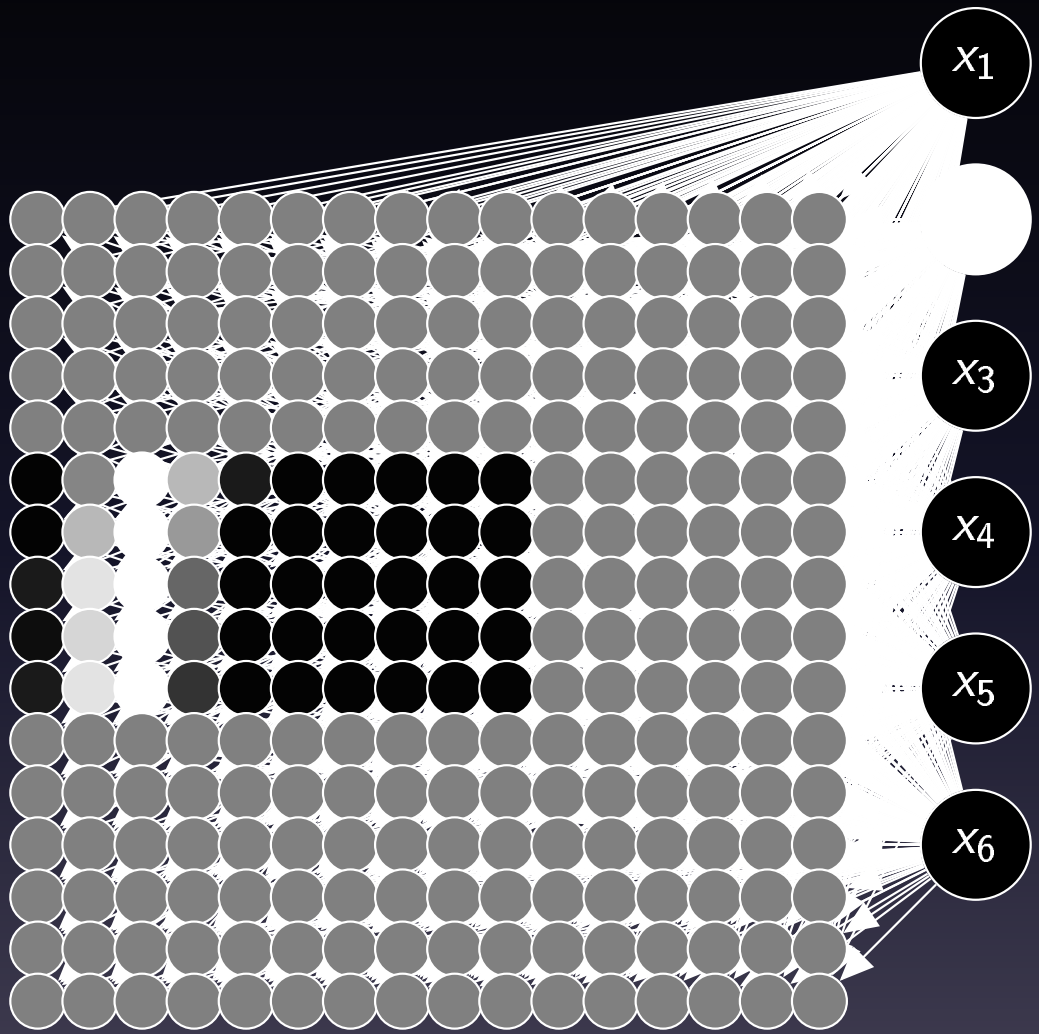

x_5
x_6

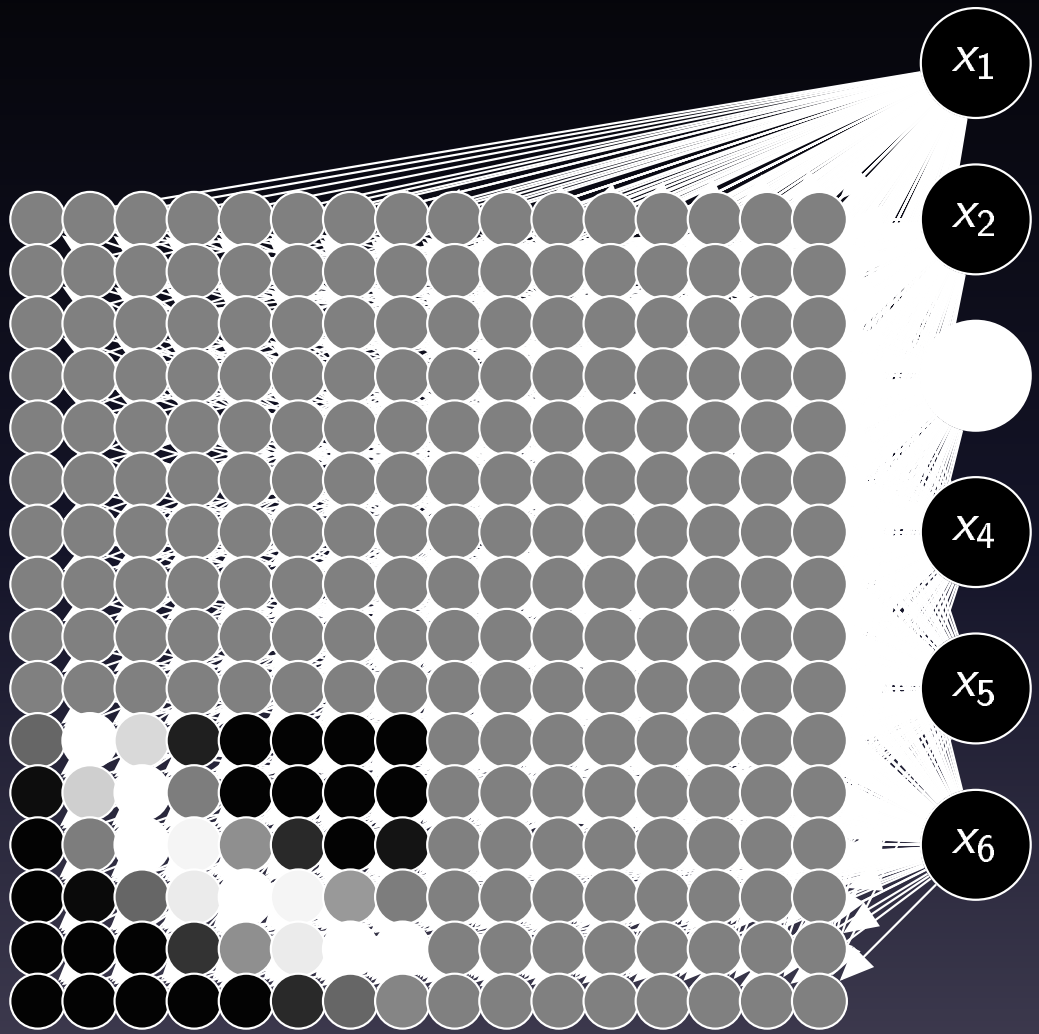

x_1
x_4
x_6

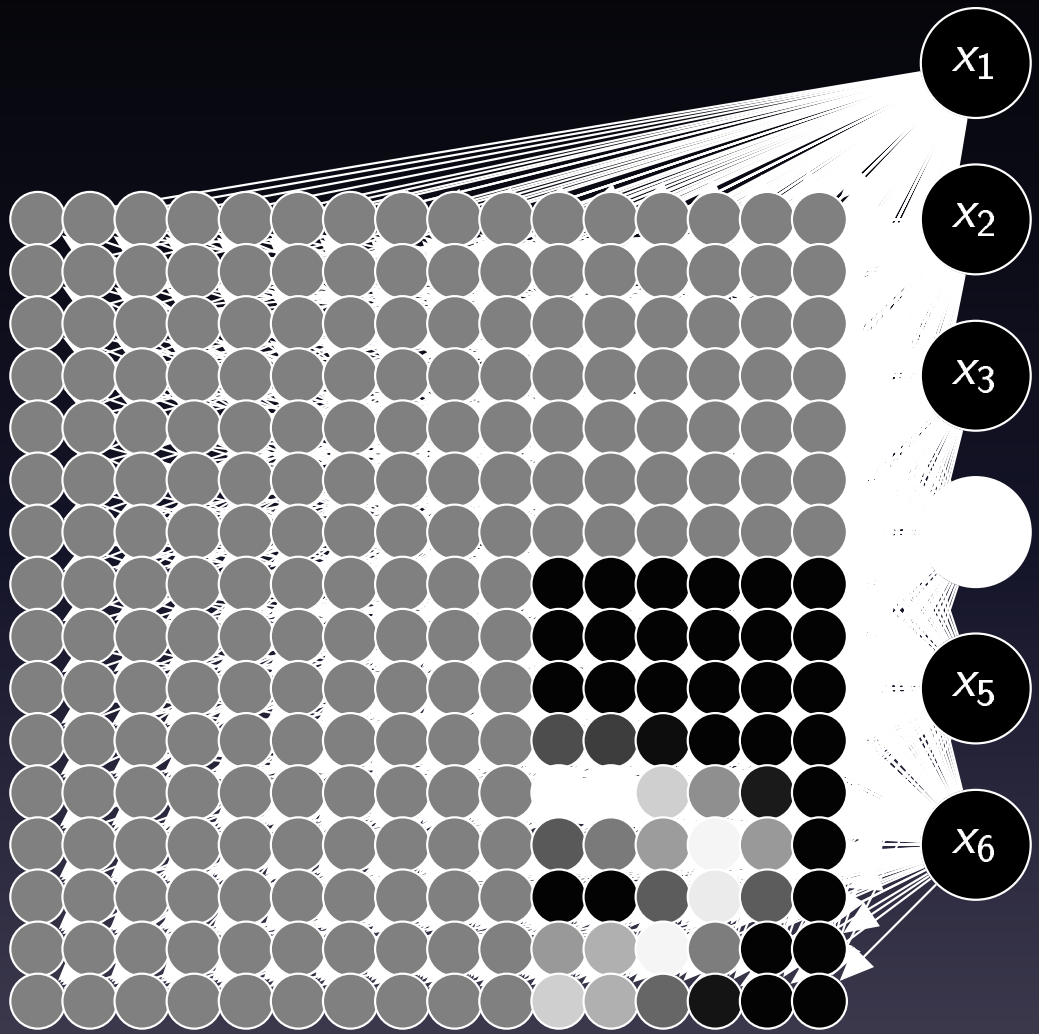
Localized Receptive Fields

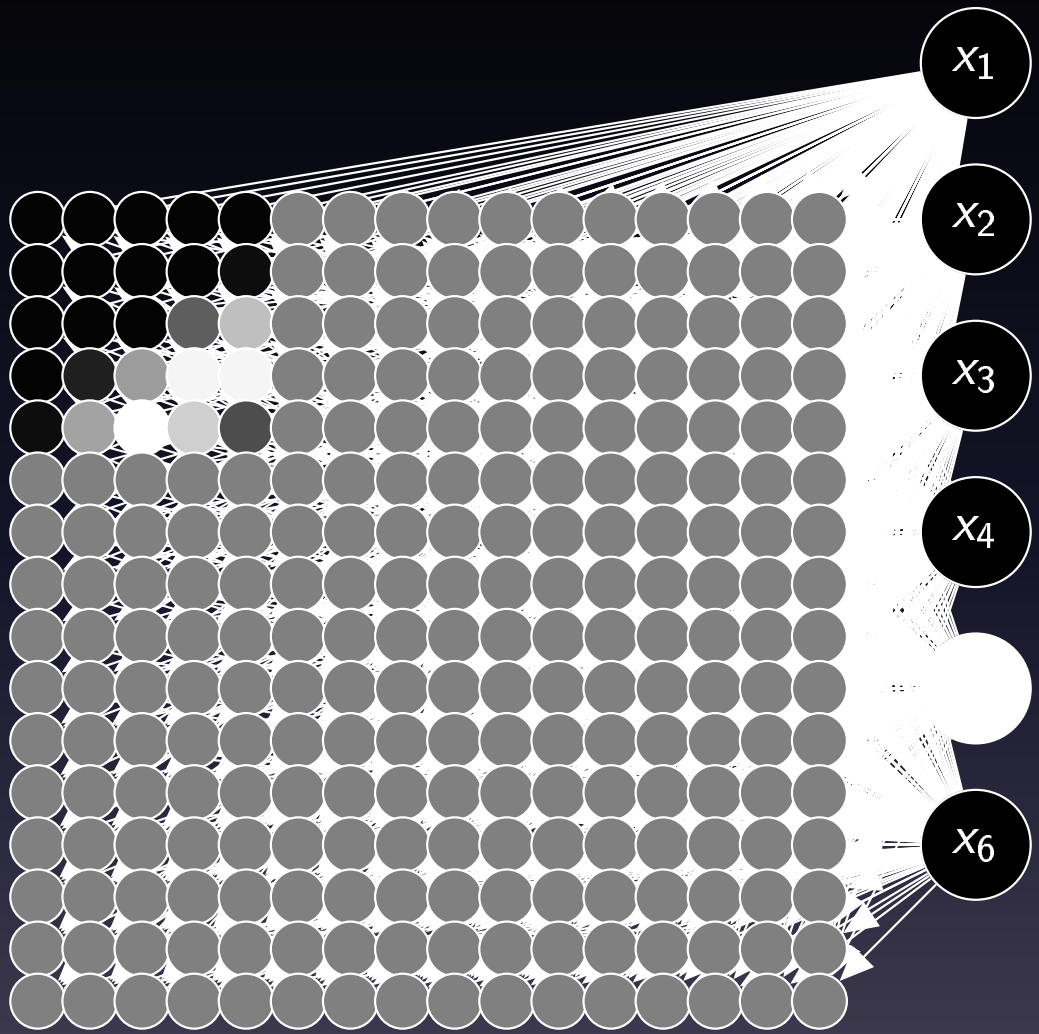
- Model can now fit global model as sum of parts.
- Each latent node associated with local features.
- Structure of model combines local features in products of experts manner (Hinton, 1999).

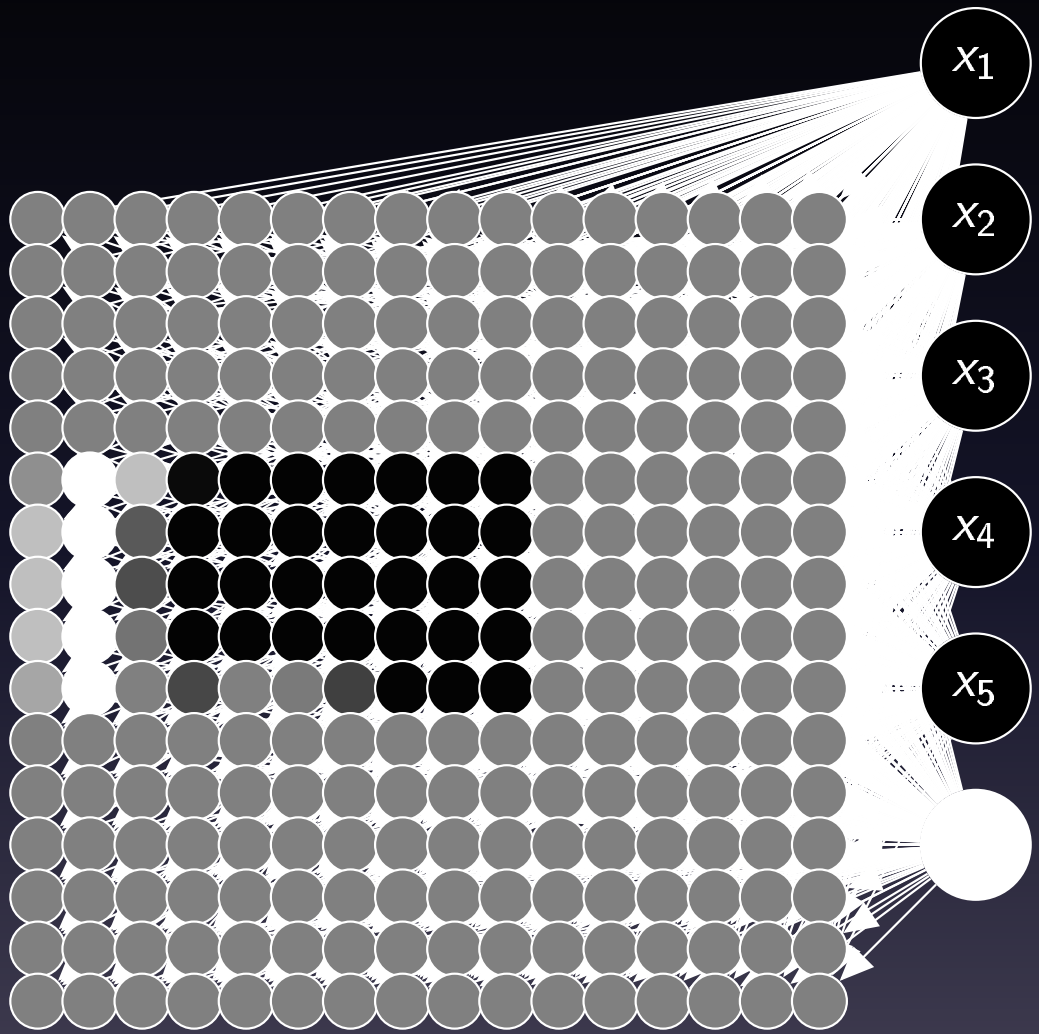

x_2
x_3
x_4
x_5
x_6

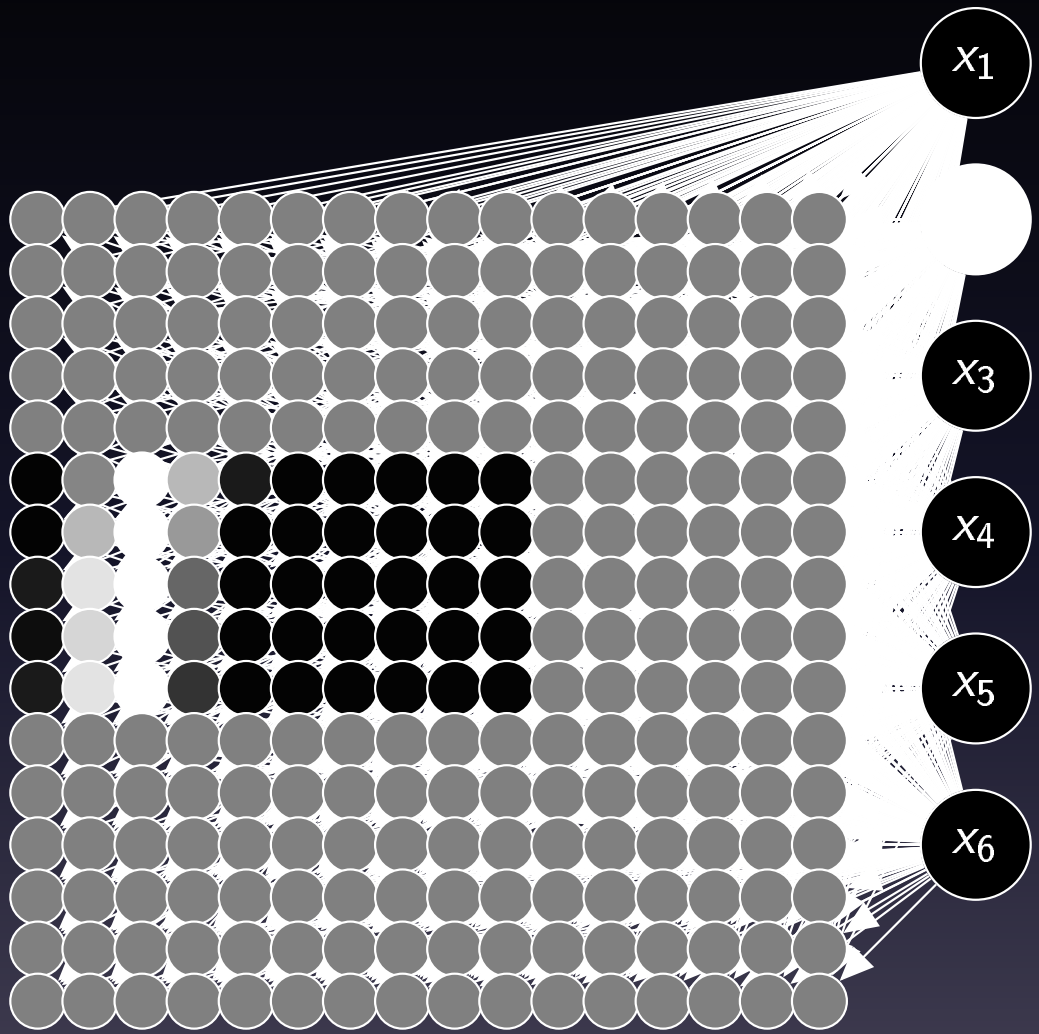

x_1
x_3
x_4
x_5
x_6

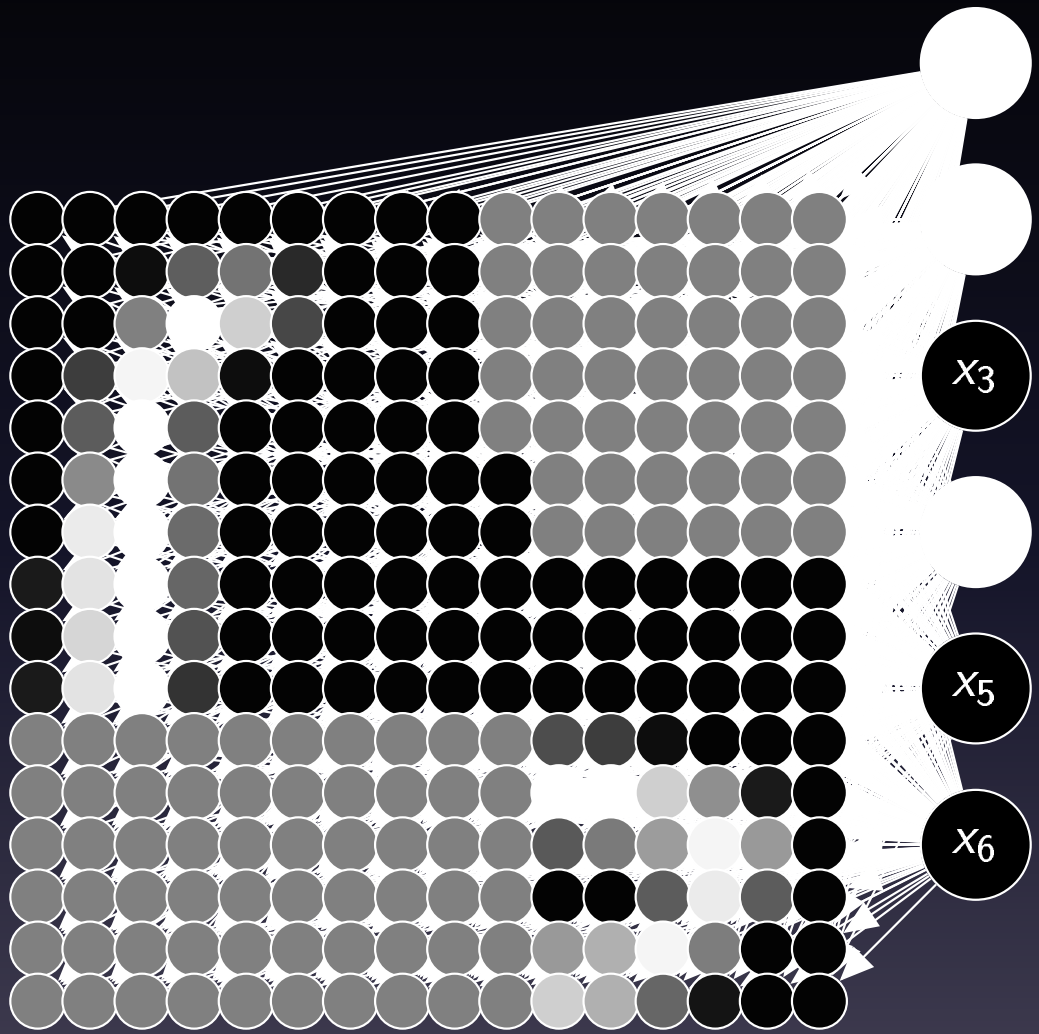

x_1
x_2
x_4
x_5
x_6

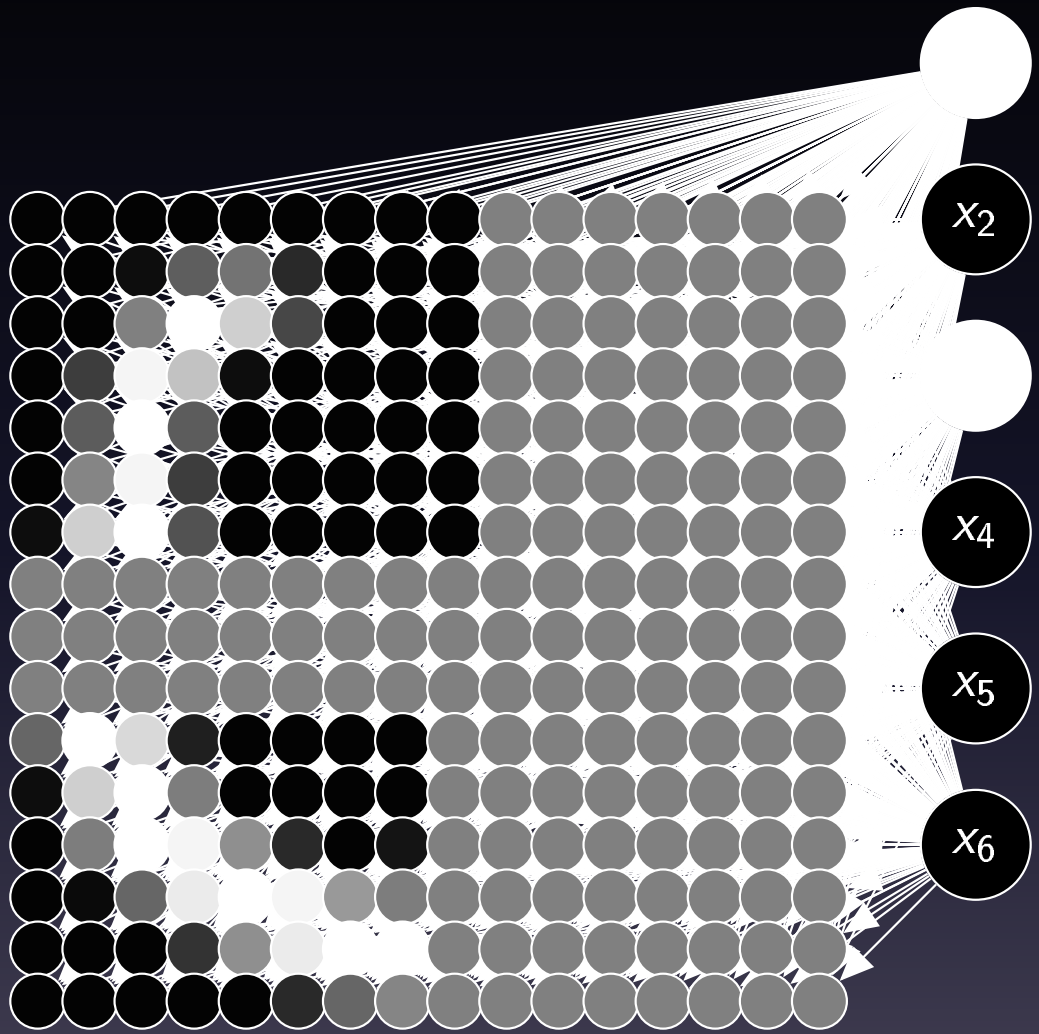

x_1
x_2
x_3
x_5
x_6

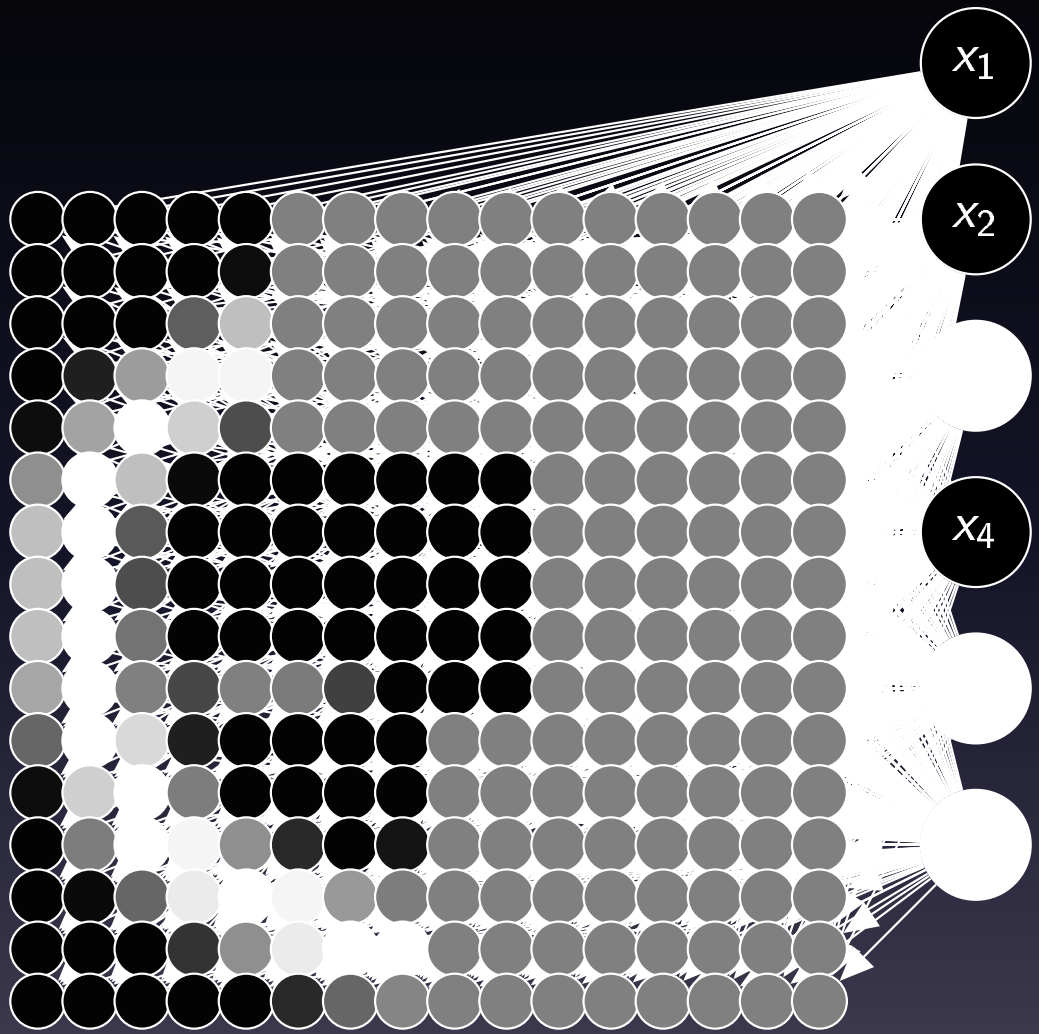

x_1
x_2
x_3
x_4
x_6

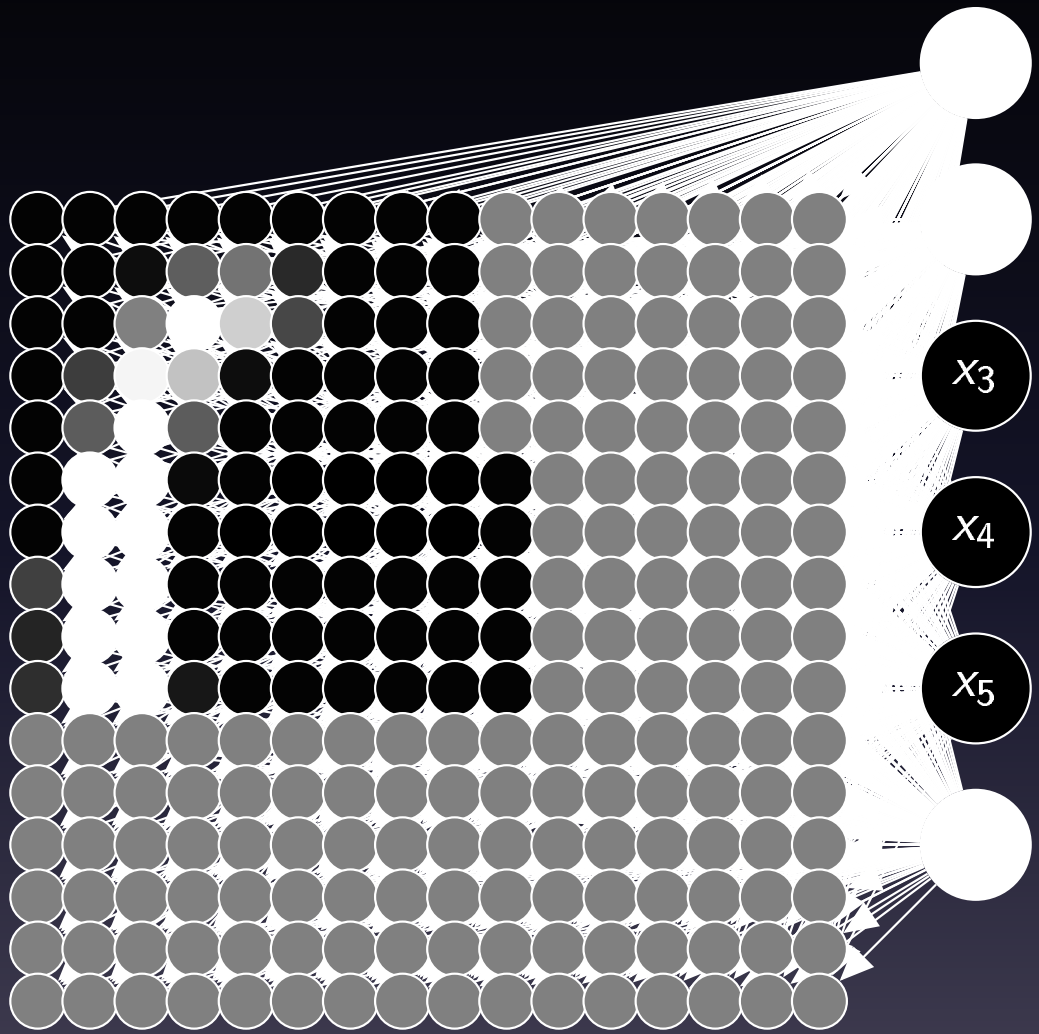

x_1
x_2
x_3
x_4
x_5

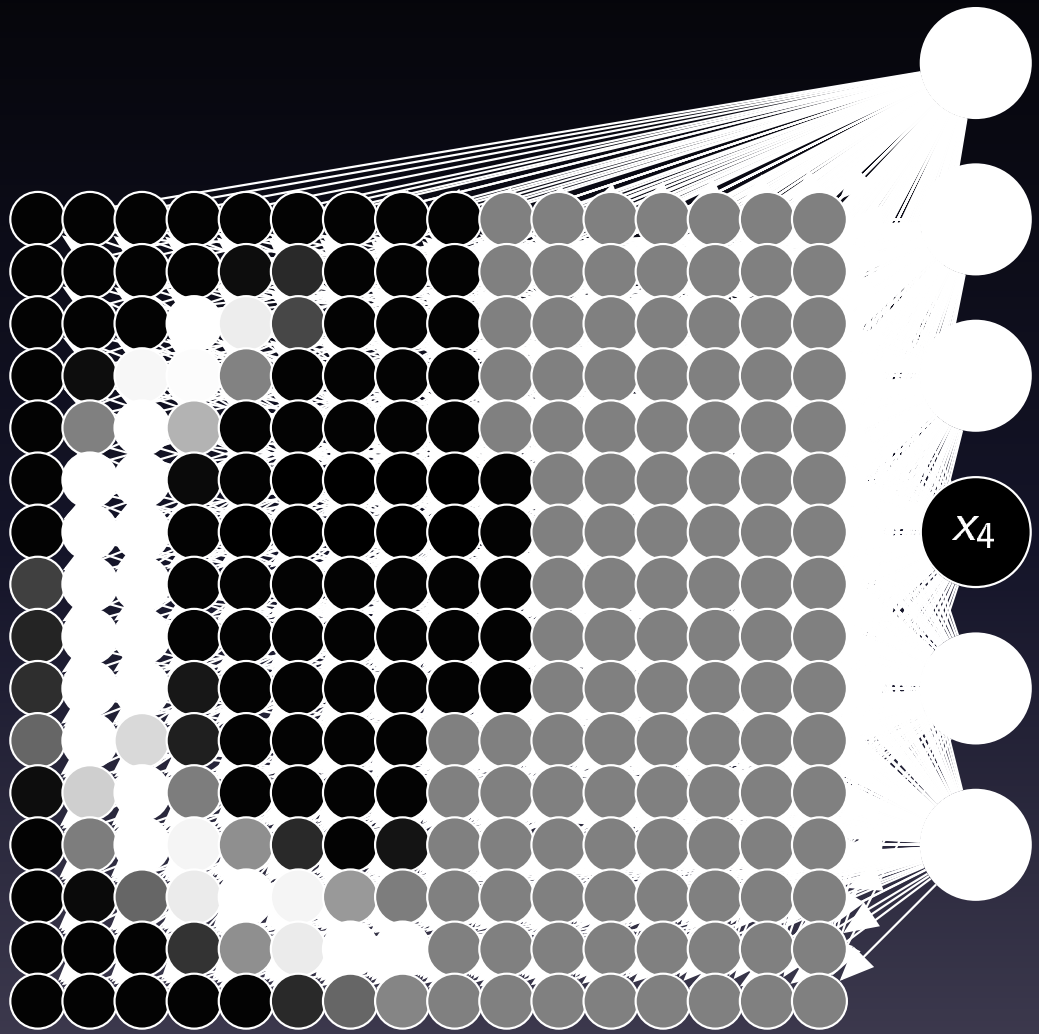

x_1
x_3
x_4
x_5
x_6


x_3
x_5
x_6


x_2
x_4
x_5
x_6


x_1
x_2
x_4

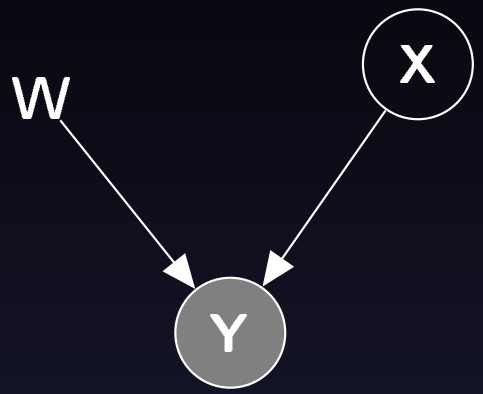

x_3
x_4
x_5


x_4

Relation between RBM and PCA/FA

- RBM is PCA with latent variables and data variables restricted binary.
- Binary restriction means latent features combine in a non-linear way.
- In PCA latent features always combine in a linear way.

PCA and RBM

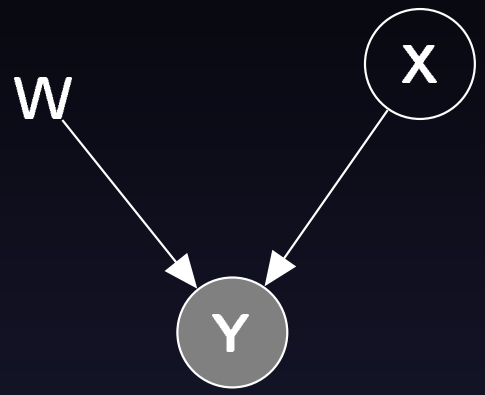


$$P\left(\mathbf{y}_{i,:}|\mathbf{x}_{i,:},\mathbf{W}\right) \propto \exp\left(-\left(\mathbf{y}_{i,:}-\mathbf{W}\mathbf{x}_{i,:}\right)^{\top}\operatorname{diag}(\mathbf{c})\left(\mathbf{y}_{i,:}-\mathbf{W}\mathbf{x}_{i,:}\right)\right)$$

$$P\left(\mathbf{x}_{i,:}\right) \propto \exp\left(\mathbf{x}_{i,:}^{\top}\operatorname{diag}\left(\mathbf{b}\right)\mathbf{x}_{i,:}\right)$$

$$P\left(\mathbf{y}_{i,:}|\mathbf{W}\right) = \sum_{\mathbf{x}_{i,:}} P\left(\mathbf{y}_{i,:}|\mathbf{x}_{i,:},\mathbf{W}\right) P\left(\mathbf{x}_{i,:}\right)$$

PCA and RBM

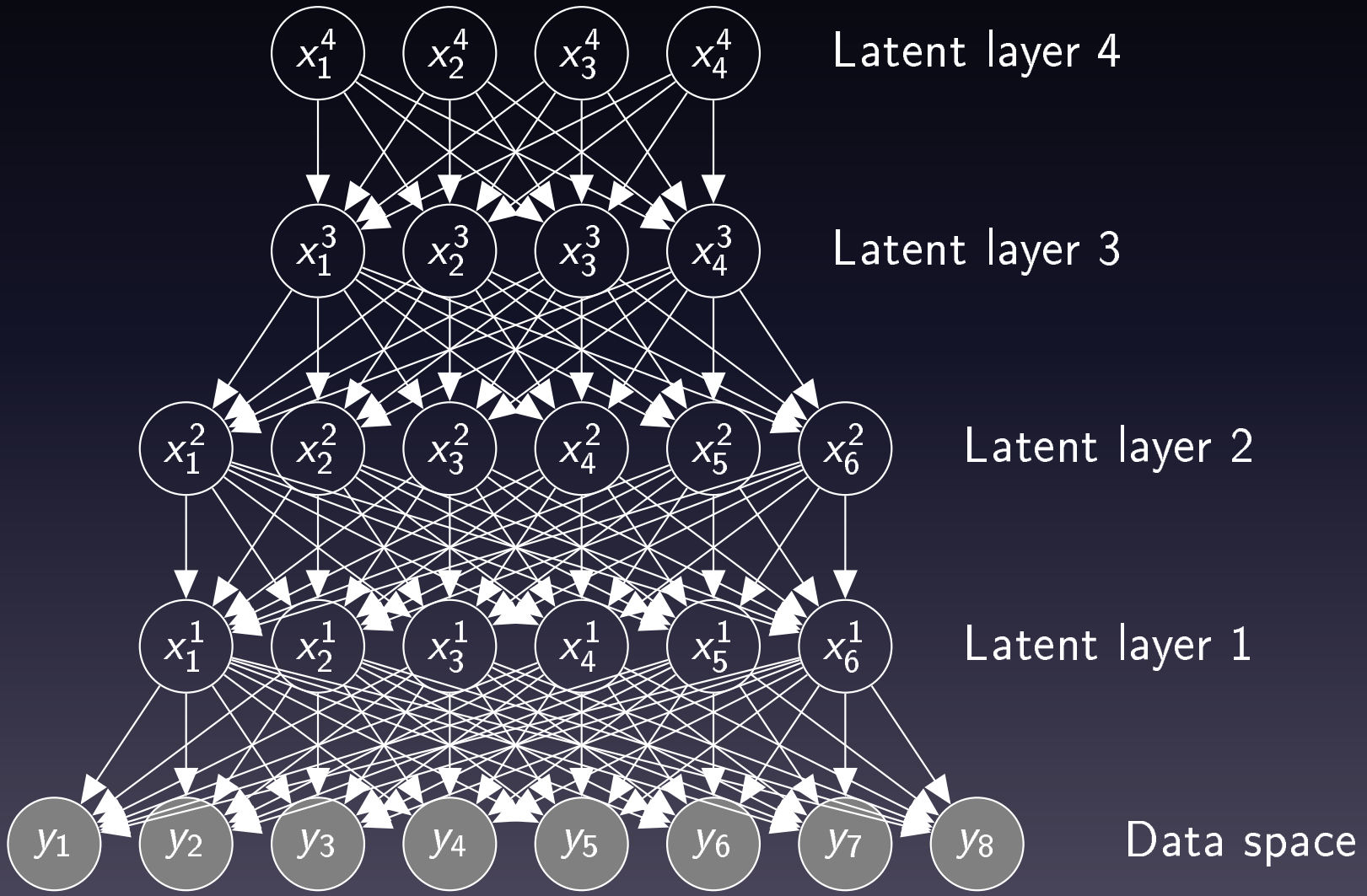


$$p\left(\mathbf{y}_{i,:}|\mathbf{x}_{i,:},\mathbf{W}\right) \propto \exp\left(-\frac{1}{2\sigma^2}\left(\mathbf{y}_{i,:}-\mathbf{W}\mathbf{x}_{i,:}\right)^\top\left(\mathbf{y}_{i,:}-\mathbf{W}\mathbf{x}_{i,:}\right)\right)$$

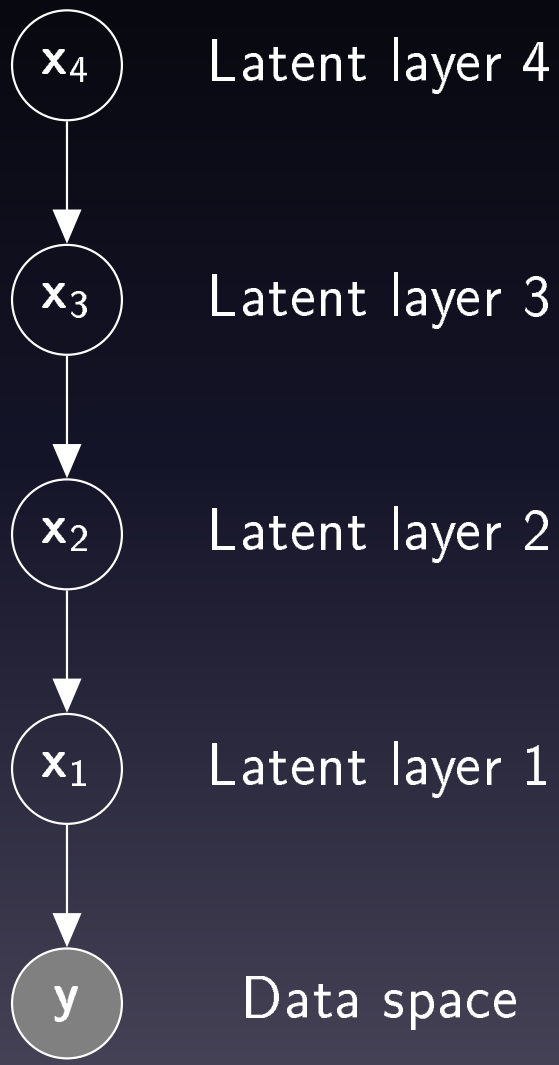
$$p\left(\mathbf{x}_{i,:}\right) \propto \exp\left(-\frac{1}{2}\mathbf{x}_{i,:}^\top\mathbf{x}_{i,:}\right)$$

$$p\left(\mathbf{y}_{i,:}|\mathbf{W}\right) = \mathcal{N}\left(\mathbf{y}_{i,:}|\mathbf{0},\mathbf{W}\mathbf{W}^\top+\sigma^2\mathbf{I}\right)$$

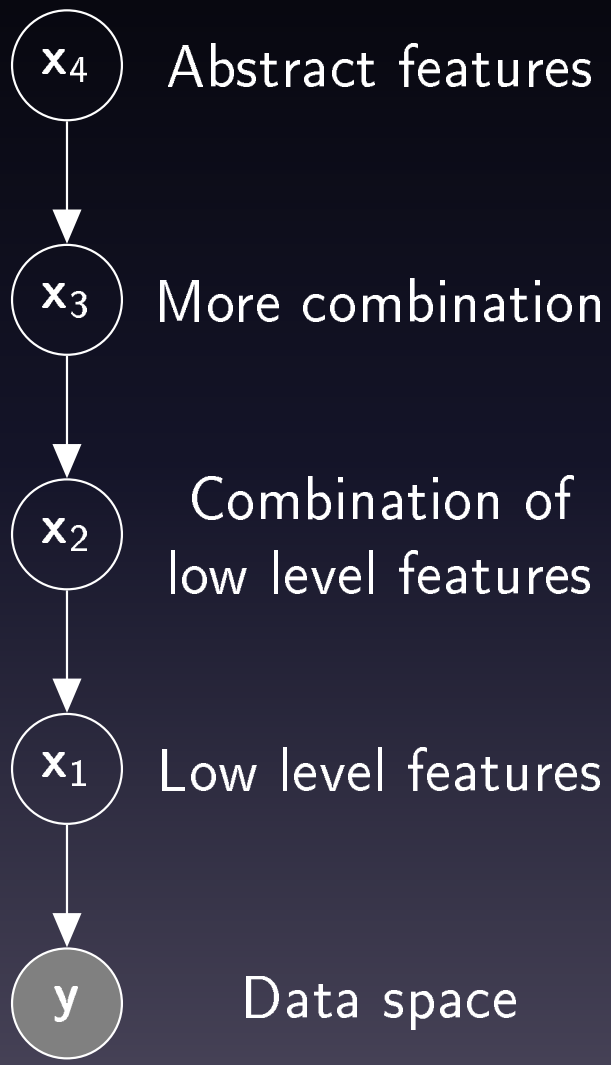
Deep Models


x_1^4 x_2^4 x_3^4 x_4^4
Latent layer 4
x_1^3 x_2^3 x_3^3 x_4^3
Latent layer 3
x_1^2 x_2^2 x_3^2 x_4^2 x_5^2 x_6^2
Latent layer 2
x_1^1 x_2^1 x_3^1 x_4^1 x_5^1 x_6^1
Latent layer 1
y_1 y_2 y_3 y_4 y_5 y_6 y_7 y_8
Data space

Deep Models


$\mathbf{x}_4$
Latent layer 4
$\mathbf{x}_3$
Latent layer 3
$\mathbf{x}_2$
Latent layer 2
$\mathbf{x}_1$
Latent layer 1
$\mathbf{y}$
Data space

Deep Models

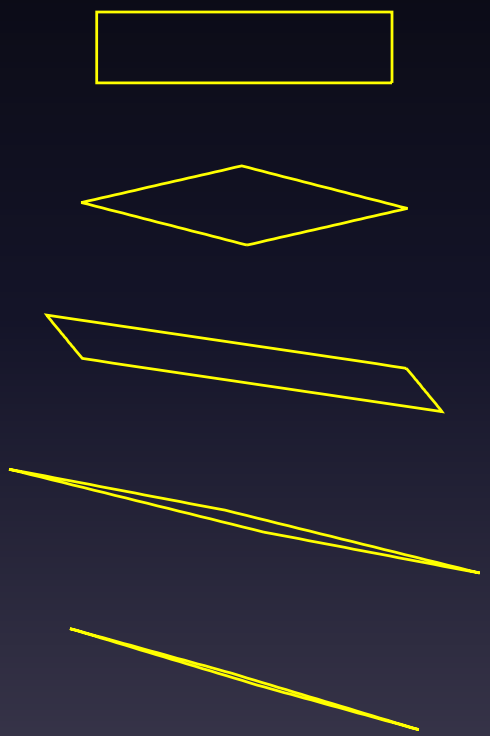

$\mathbf{x}_4$
Abstract features
$\mathbf{x}_3$
More combination
$\mathbf{x}_2$
Combination of
low level features
$\mathbf{x}_1$
Low level features
$\mathbf{y}$
Data space

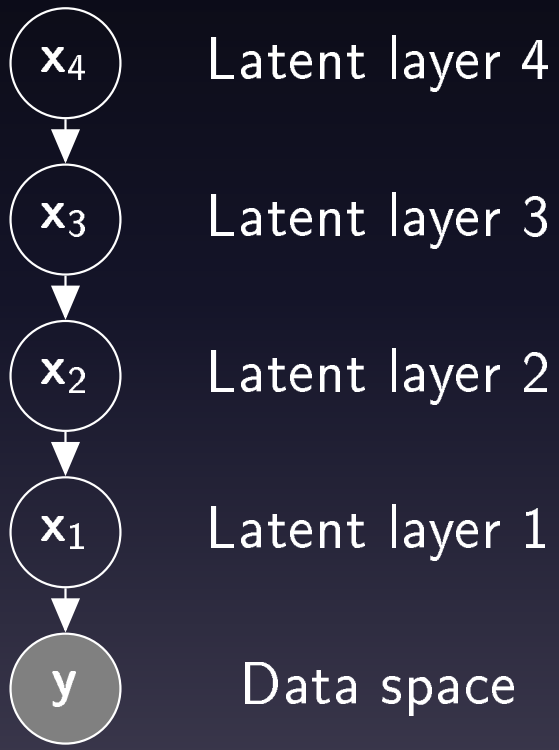
Deep Gaussian Processes

(Damianou and Lawrence, 2013)

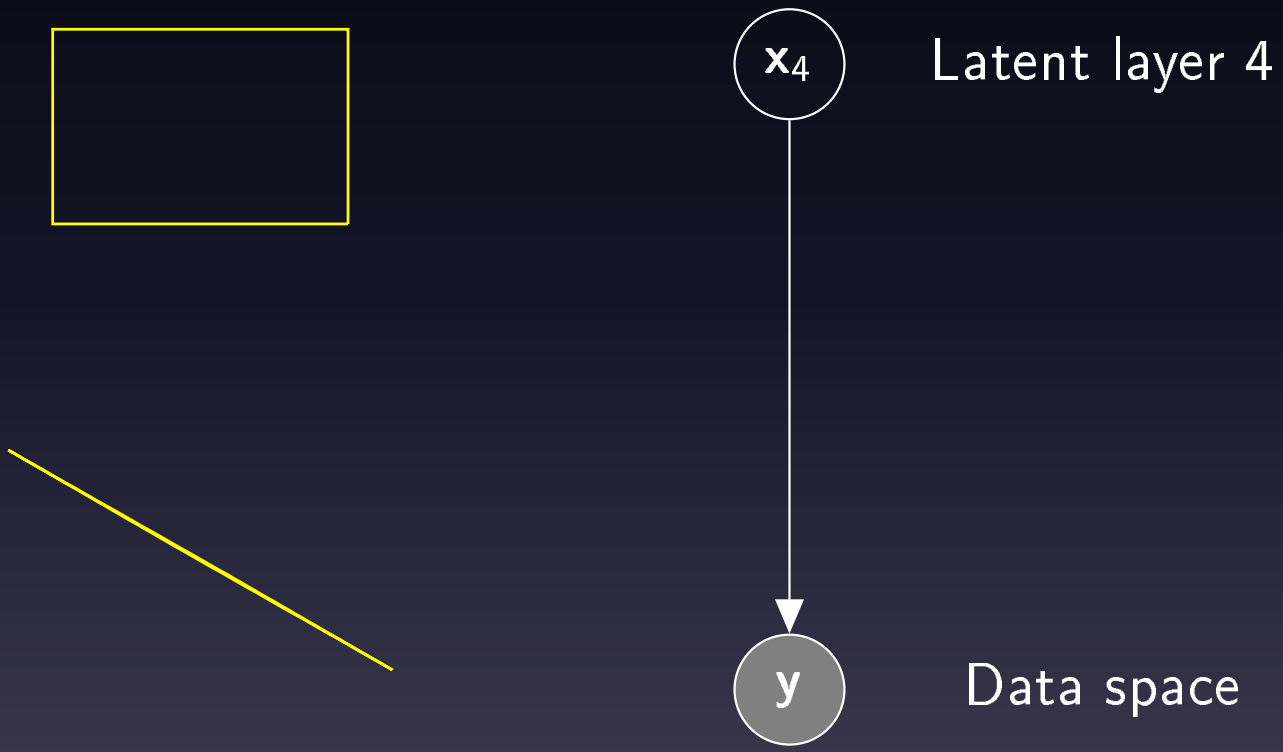
- Deep architectures allow abstraction of features (Bengio, 2009; Hinton and Osindero, 2006; Salakhutdinov and Murray, 2008).
- We use variational approach to stack GP models.

Stacked PCA

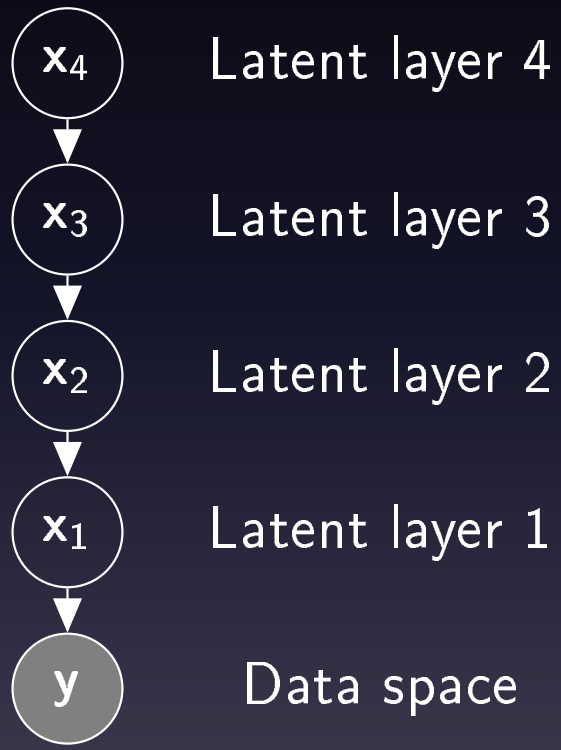



x_4
Latent layer 4
x_3
Latent layer 3
x_2
Latent layer 2
x_1
Latent layer 1
y
Data space

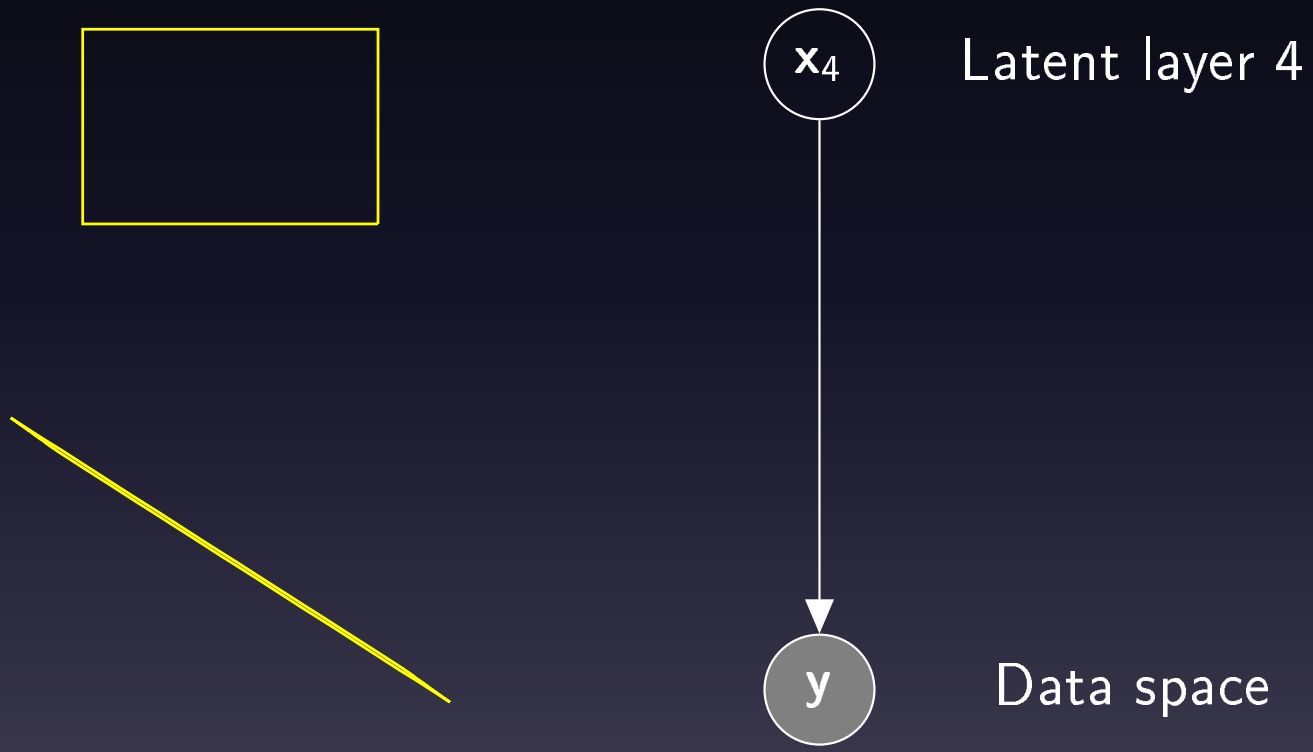
Stacked PCA



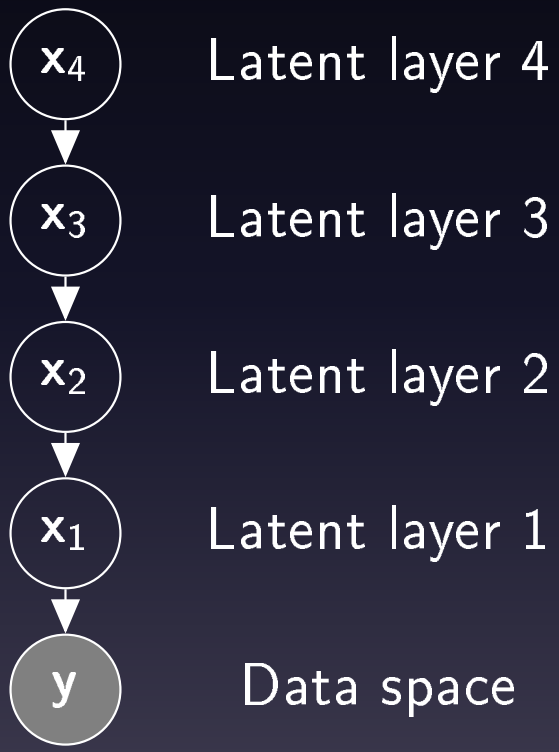
Stacked PCA


$\mathbf{x}_4$
Latent layer 4
$\mathbf{x}_3$
Latent layer 3
$\mathbf{x}_2$
Latent layer 2
$\mathbf{x}_1$
Latent layer 1
$\mathbf{y}$
Data space

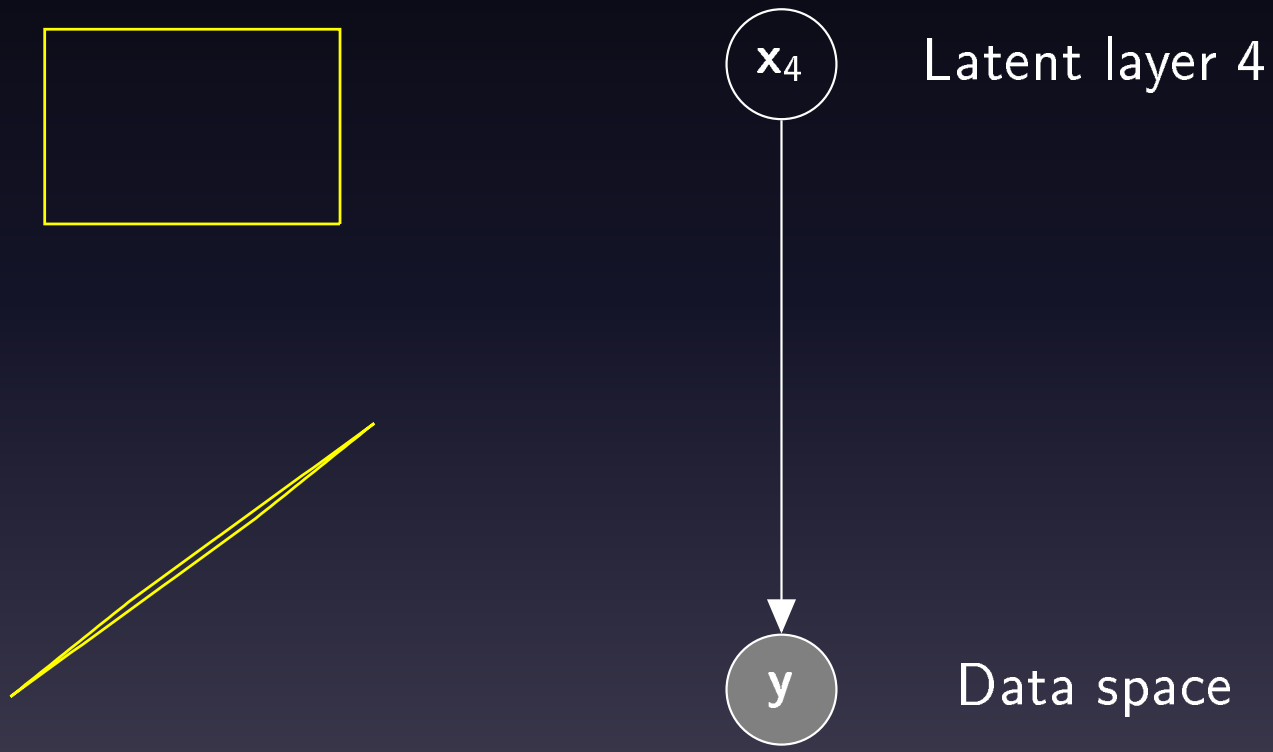
Stacked PCA



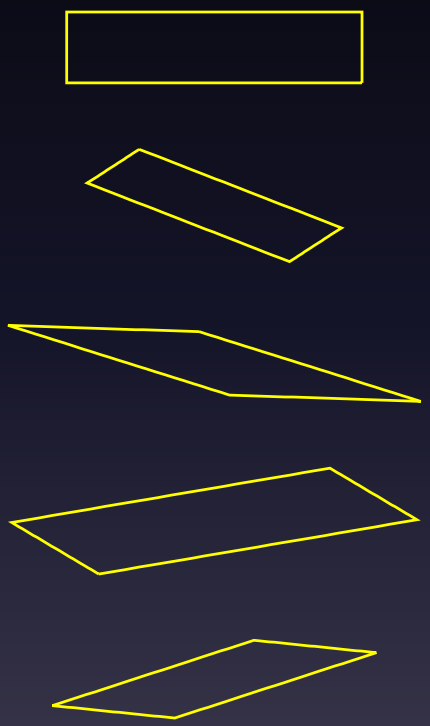
Stacked PCA

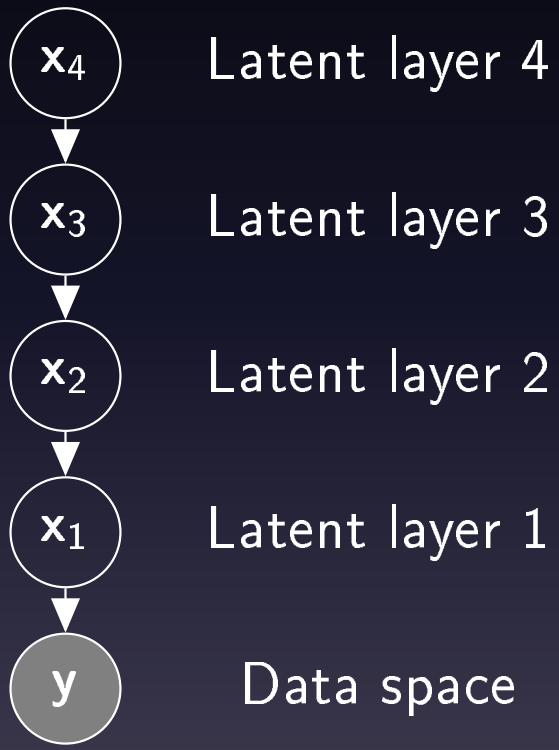

x_4
Latent layer 4
x_3
Latent layer 3
x_2
Latent layer 2
x_1
Latent layer 1
y
Data space

Stacked PCA

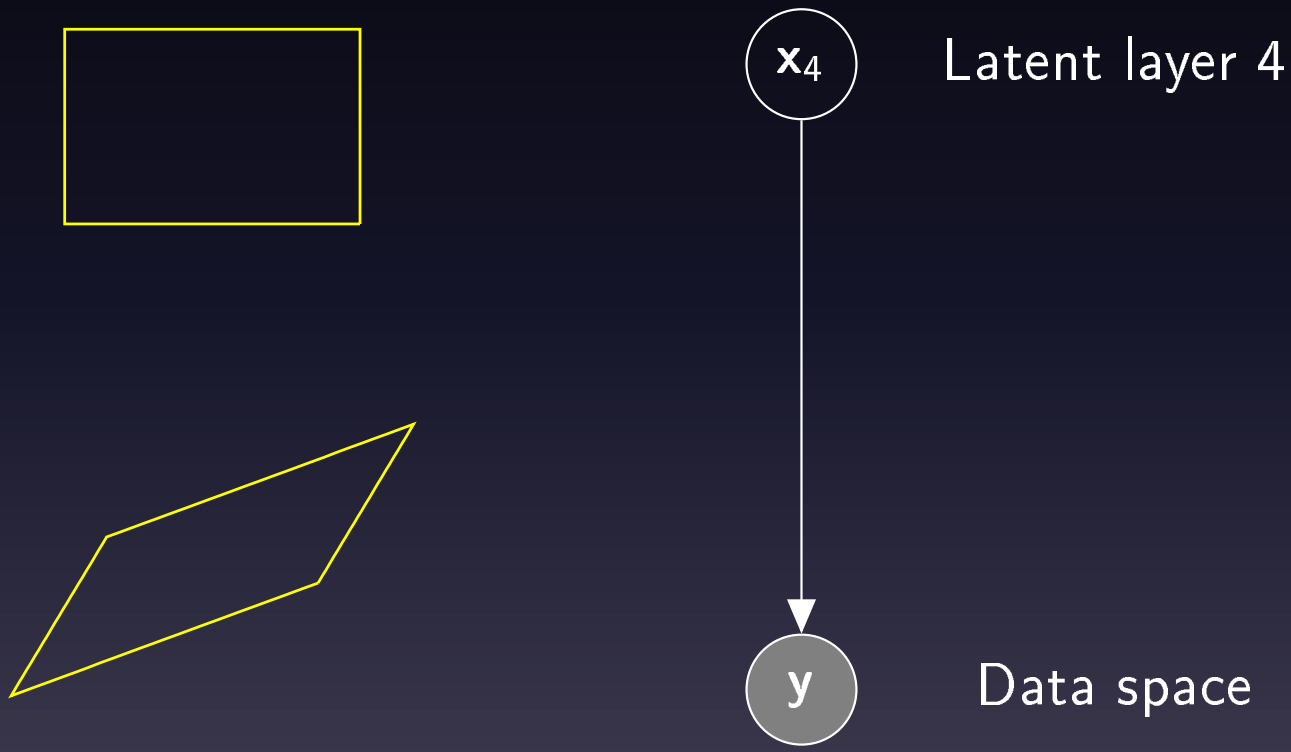


Stacked PCA

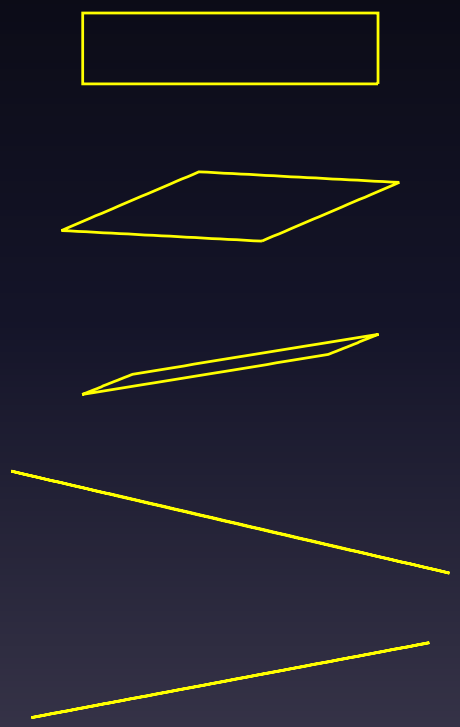


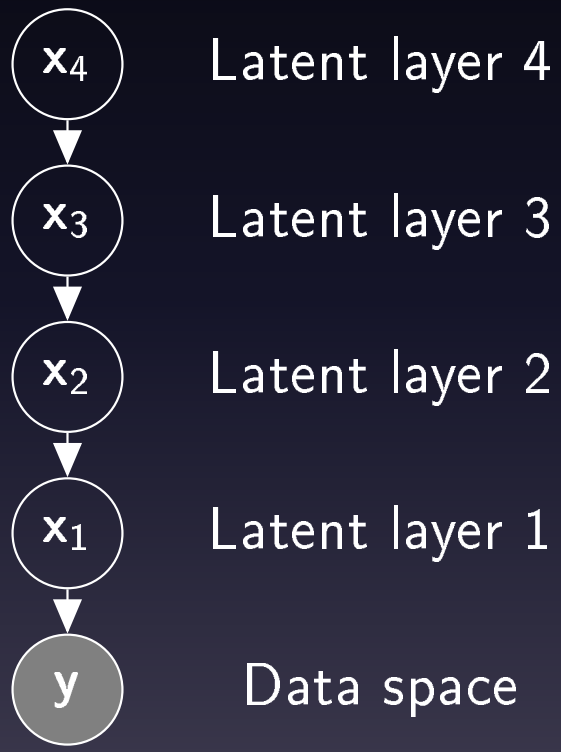

$\mathbf{x}_4$
Latent layer 4
$\mathbf{x}_3$
Latent layer 3
$\mathbf{x}_2$
Latent layer 2
$\mathbf{x}_1$
Latent layer 1
$\mathbf{y}$
Data space

Stacked PCA

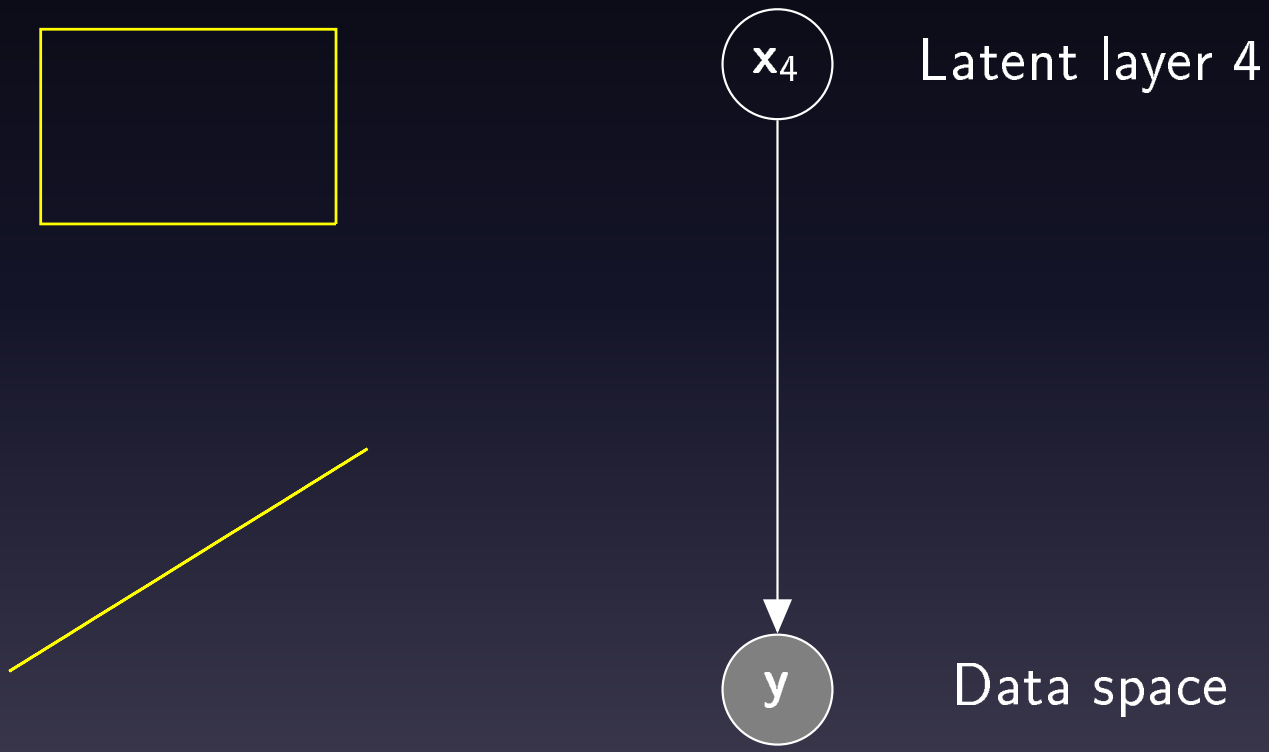


Stacked PCA

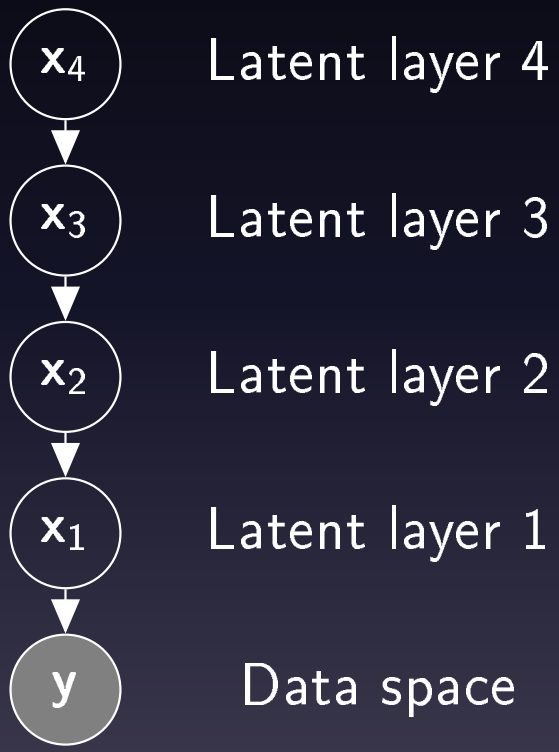



$\mathbf{x}_4$
Latent layer 4
$\mathbf{x}_3$
Latent layer 3
$\mathbf{x}_2$
Latent layer 2
$\mathbf{x}_1$
Latent layer 1
$\mathbf{y}$
Data space

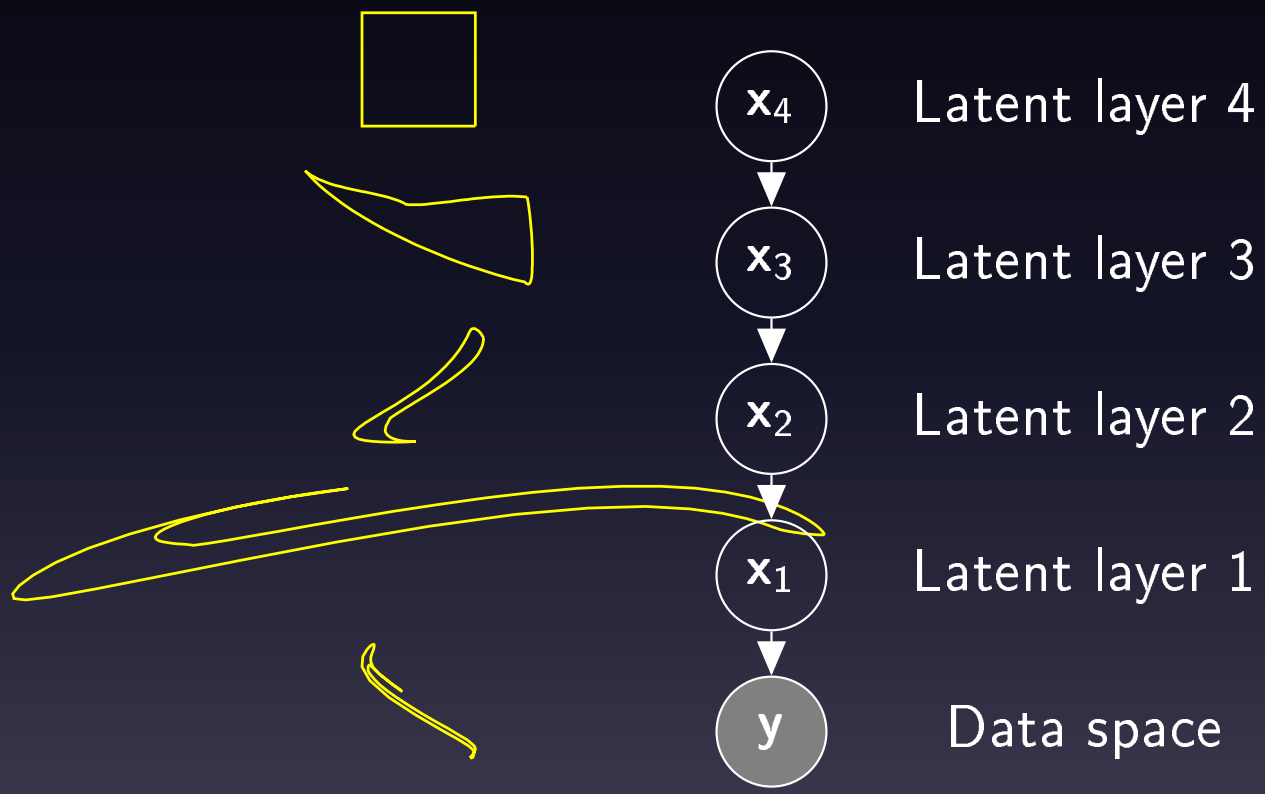
Stacked PCA



Stacked GPs


$\mathbf{x}_4$
Latent layer 4
$\mathbf{x}_3$
Latent layer 3
$\mathbf{x}_2$
Latent layer 2
$\mathbf{x}_1$
Latent layer 1
$\mathbf{y}$
Data space

Stacked GPs


$\mathbf{x}_4$
Latent layer 4
$\mathbf{x}_3$
Latent layer 3
$\mathbf{x}_2$
Latent layer 2
$\mathbf{x}_1$
Latent layer 1
$\mathbf{y}$
Data space

Stacked GPs

Latent layer 4

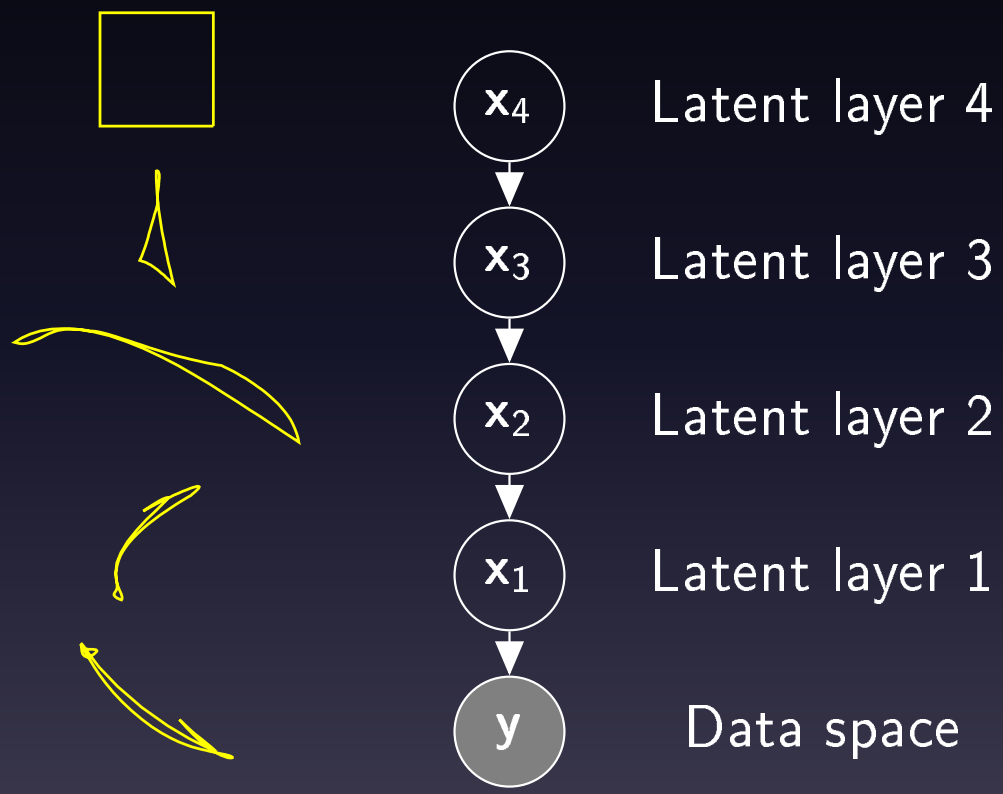
Latent layer 3

Latent layer 2

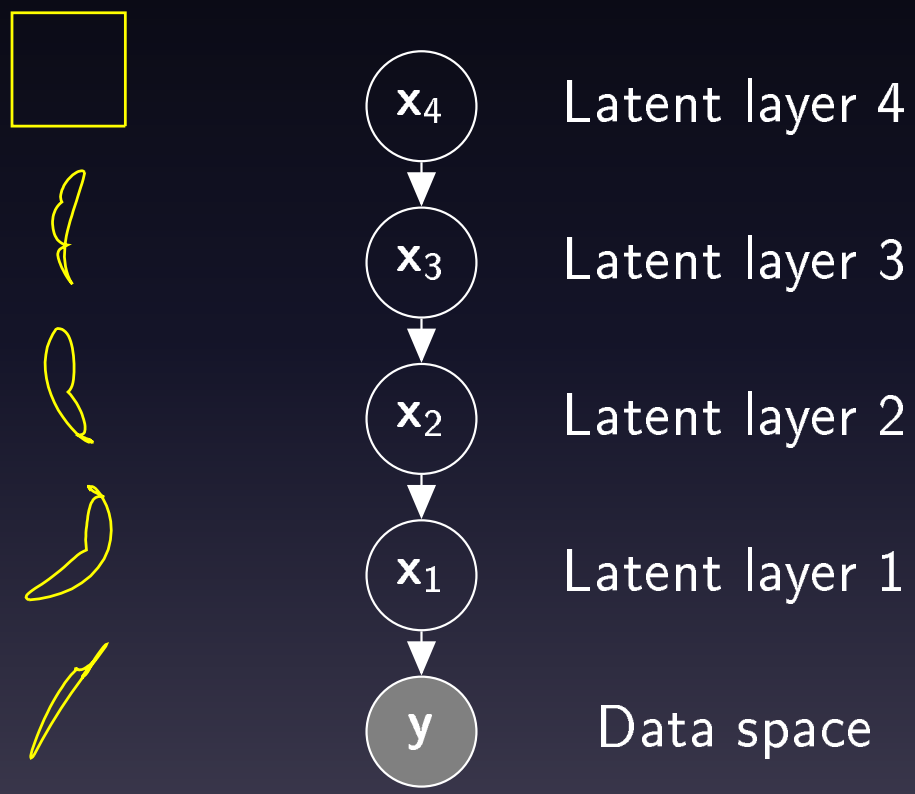
Latent layer 1

Data space

Stacked GPs


$\mathbf{x}_4$
Latent layer 4
$\mathbf{x}_3$
Latent layer 3
$\mathbf{x}_2$
Latent layer 2
$\mathbf{x}_1$
Latent layer 1
$\mathbf{y}$
Data space

Stacked GPs


$\mathbf{x}_4$
Latent layer 4
$\mathbf{x}_3$
Latent layer 3
$\mathbf{x}_2$
Latent layer 2
$\mathbf{x}_1$
Latent layer 1
$\mathbf{y}$
Data space

Deep GPs

- Stacking PPCA still leads to a linear latent variable model.
- To stack latent variable models, need a non-linear model.
- The GP-LVM is a non-linear latent variable model.
- Stacking GP-LVM leads to hierarchical GP-LVM.

Hierarchical GP-LVM

(Lawrence and Moore, 2007)

Stacking Gaussian Processes

- Regressive dynamics provides a simple hierarchy.
    - The input space of the GP is governed by another GP.
- By stacking GPs we can consider more complex hierarchies.
- Ideally we should marginalise latent spaces
    - In practice we seek MAP solutions.

Two Correlated Subjects

(Lawrence and Moore, 2007)

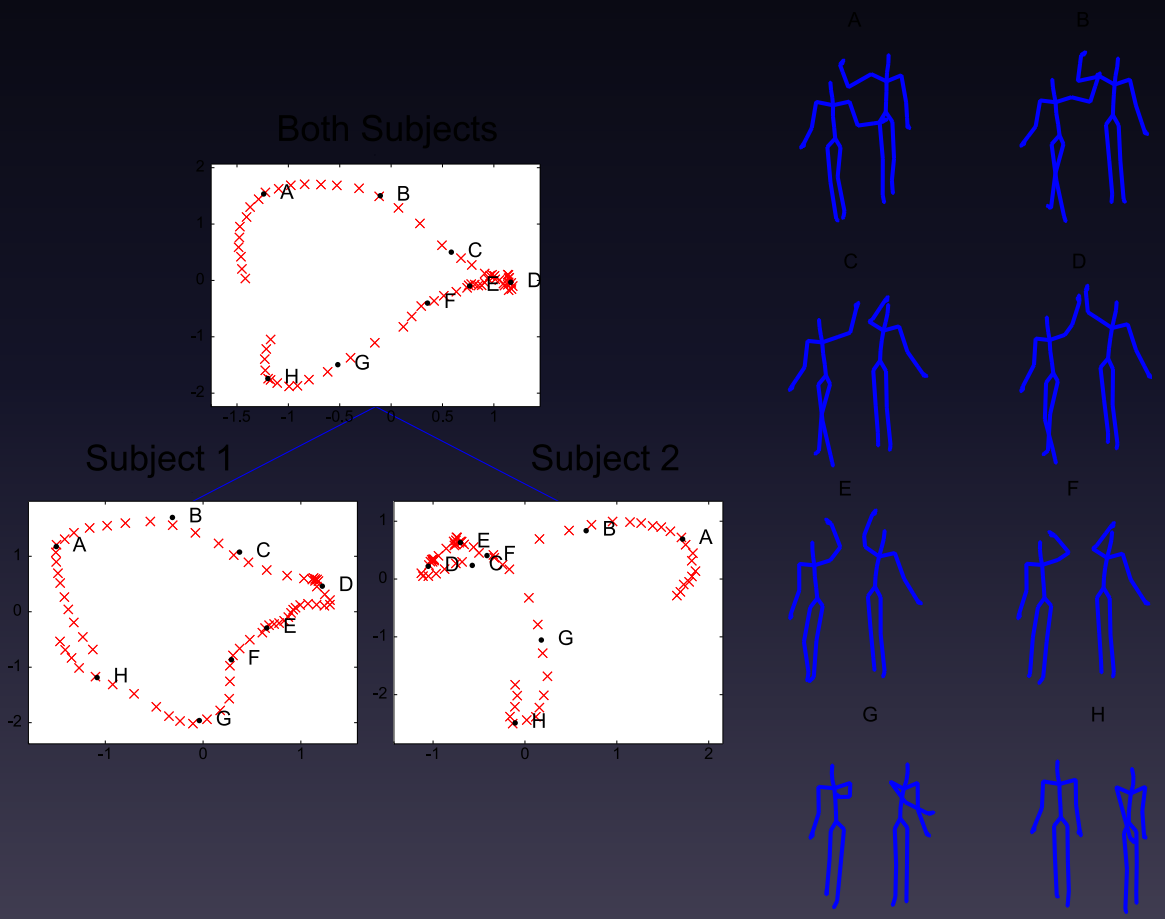


Figure: Hierarchical model of a 'high five'.

Within Subject Hierarchy

(Lawrence and Moore, 2007)

Decomposition of Body

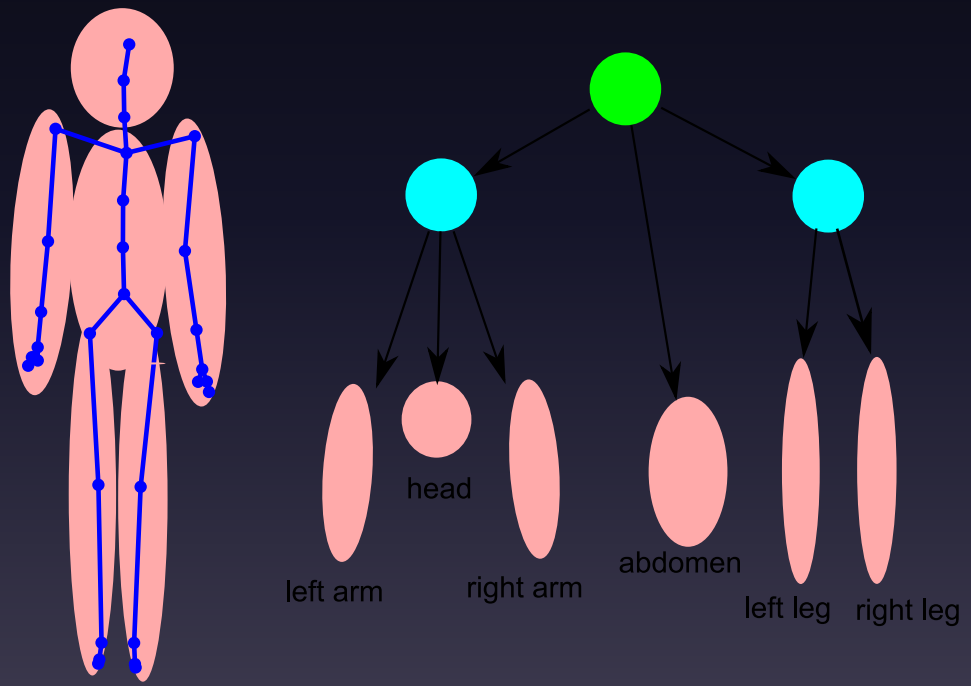


Figure: Decomposition of a subject.

Single Subject Run/Walk

(Lawrence and Moore, 2007)

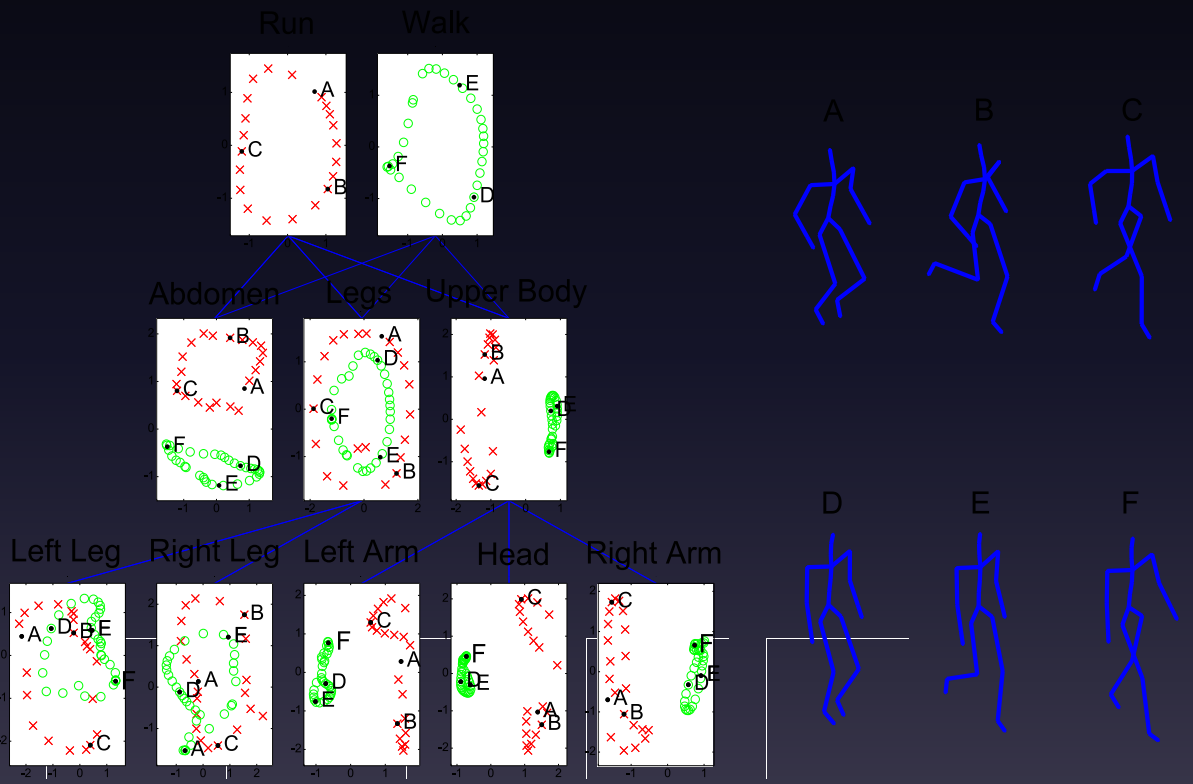
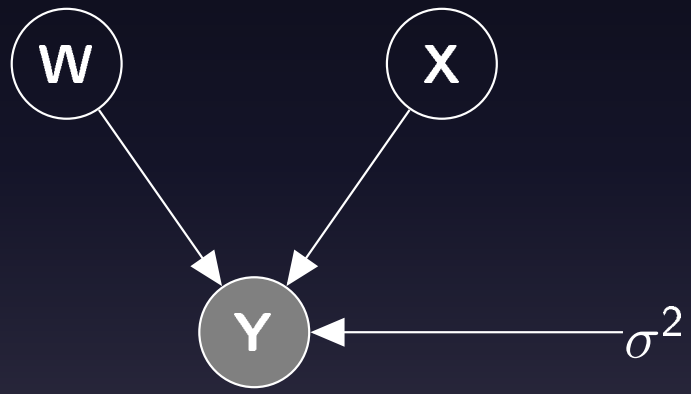


Figure: Hierarchical model of a walk and a run.

▸ Return

Bayesian GP-LVM

- Bayesian GP-LVM allows variational marginalization of $\mathbf{X}$ and $\mathbf{W}$.



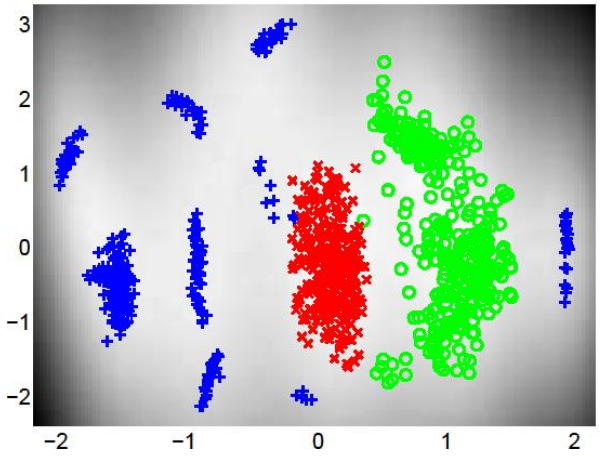
- This leads to a Bayesian model where latent dimensionality can be learnt.

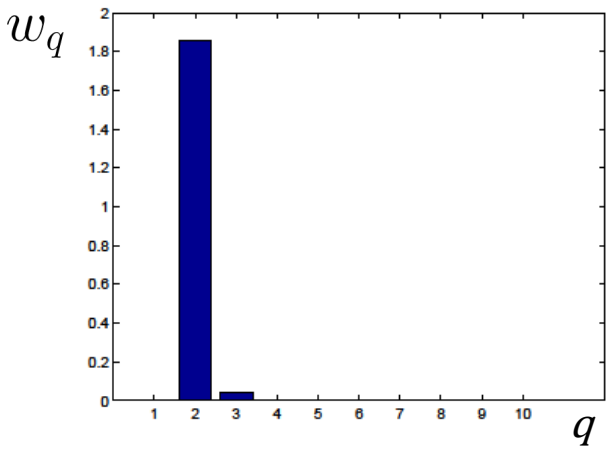
Automatic dimensionality detection

- Achieved by employing an *Automatic Relevance Determination (ARD)* covariance function for the prior on the GP mapping
- $f \sim GP(\mathbf{0}, k_f)$ with

$$k_f(\mathbf{x}_i, \mathbf{x}_j) = \sigma^2 \exp\left(-\frac{1}{2}\sum_{q=1}^{Q} w_q (x_{i,q} - x_{j,q})^2\right)$$

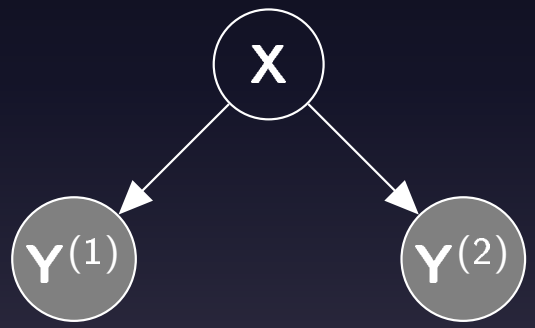
- Example





Modeling Multiple 'Views'

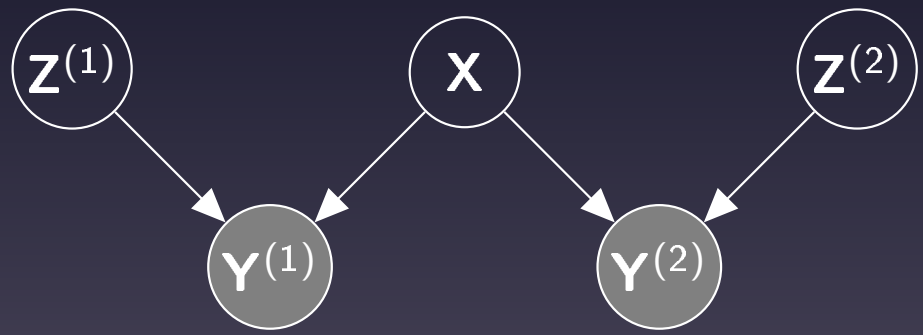
- Single space to model correlations between two different data sources, e.g., images & text, image & pose.
- Shared latent spaces: (Shon et al., 2006; Navaratnam et al., 2007; Ek et al., 2008b)



- Effective when the 'views' are correlated.
- But not all information is shared between both 'views'.
- PCA applied to concatenated data vs CCA applied to data.

Shared-Private Factorization

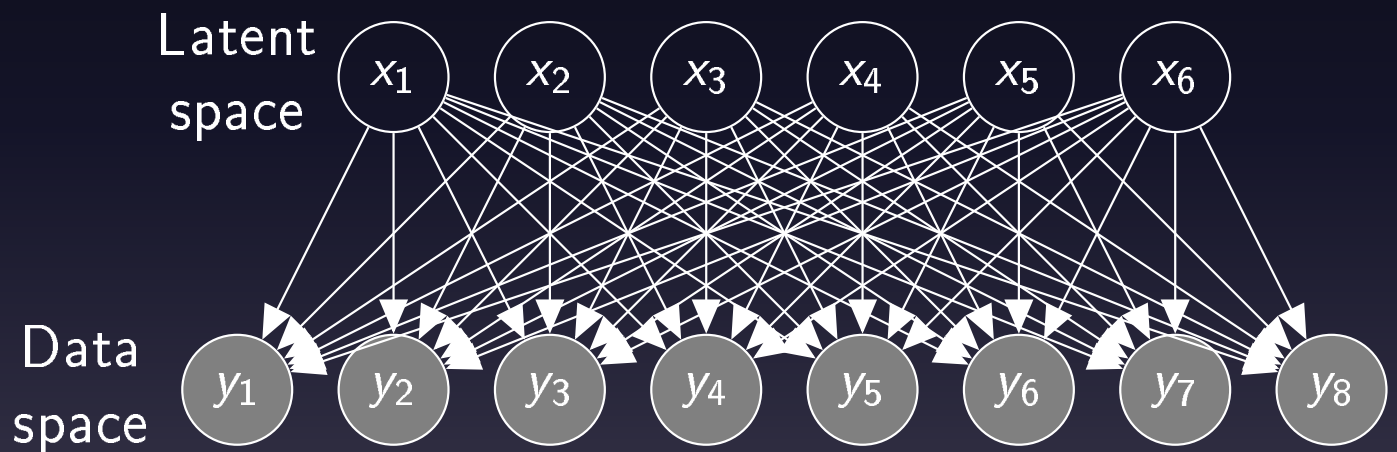
- In real scenarios, the 'views' are neither fully independent, nor fully correlated.
- Shared models
  - either allow information relevant to a single view to be mixed in the shared signal,
  - or are unable to model such private information.
- Solution: Model shared and private information (Virtanen et al., 2011; Ek et al., 2008a; Leen and Fyfe, 2006; Klami and Kaski, 2007, 2008; Tucker, 1958)



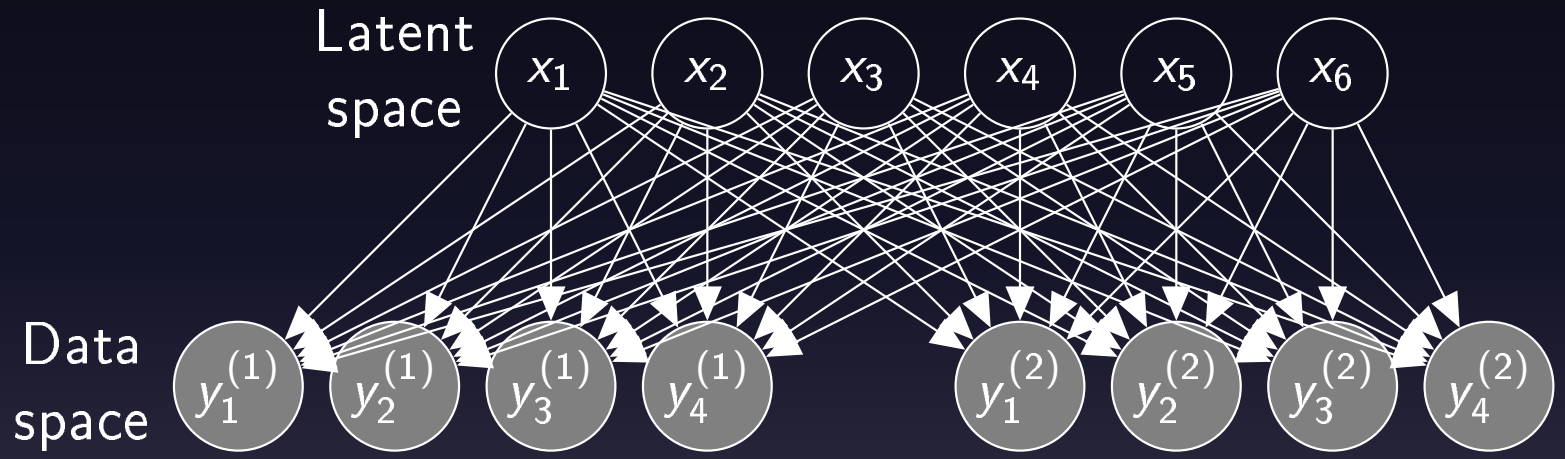
- Probabilistic CCA is case when dimensionality of $\mathbf{Z}$ matches $\mathbf{Y}^{(i)}$ (cf Inter Battery Factor Analysis (Tucker, 1958)).

Manifold Relevance Determination

(Damianou et al., 2012)



Shared GP-LVM



Separate ARD parameters for mappings to $\mathbf{Y}^{(1)}$ and $\mathbf{Y}^{(2)}$.

Example: Yale faces

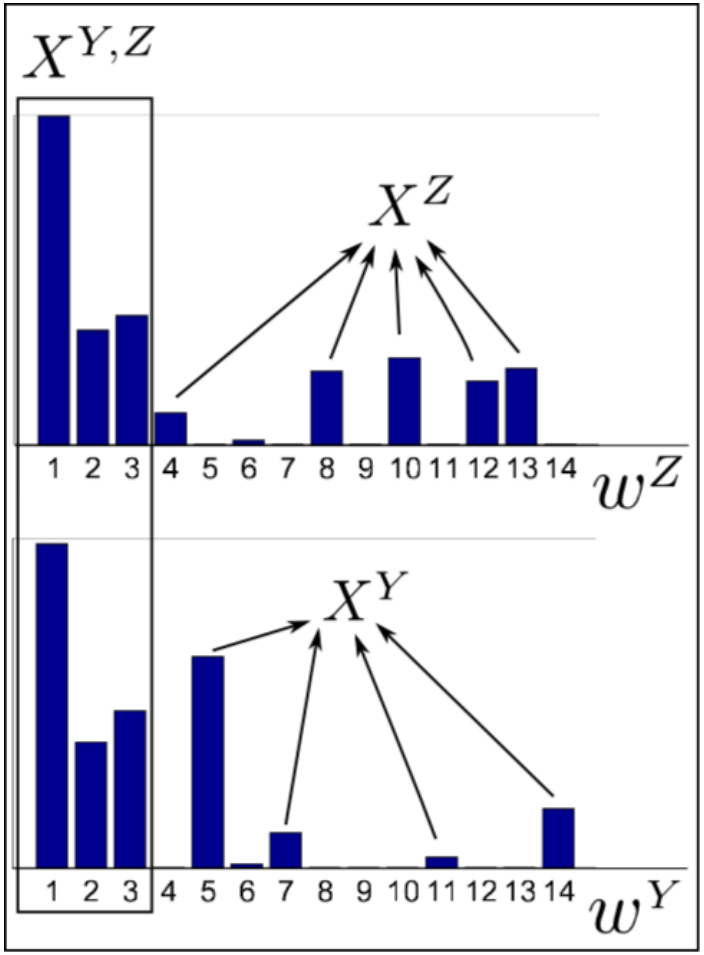
- Dataset Y: 3 persons under all illumination conditions
- Dataset Z: As above for 3 different persons
- Align datapoints $\mathbf{x}_n$ and $\mathbf{z}_n$ only based on the lighting direction

Results

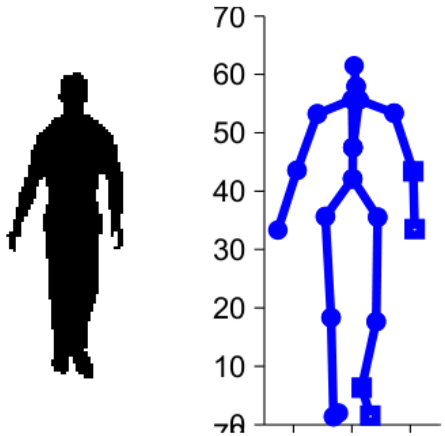
- Latent space X initialised with 14 dimensions

- Weights define a segmentation of X

- Video / demo…



Potential applications..?

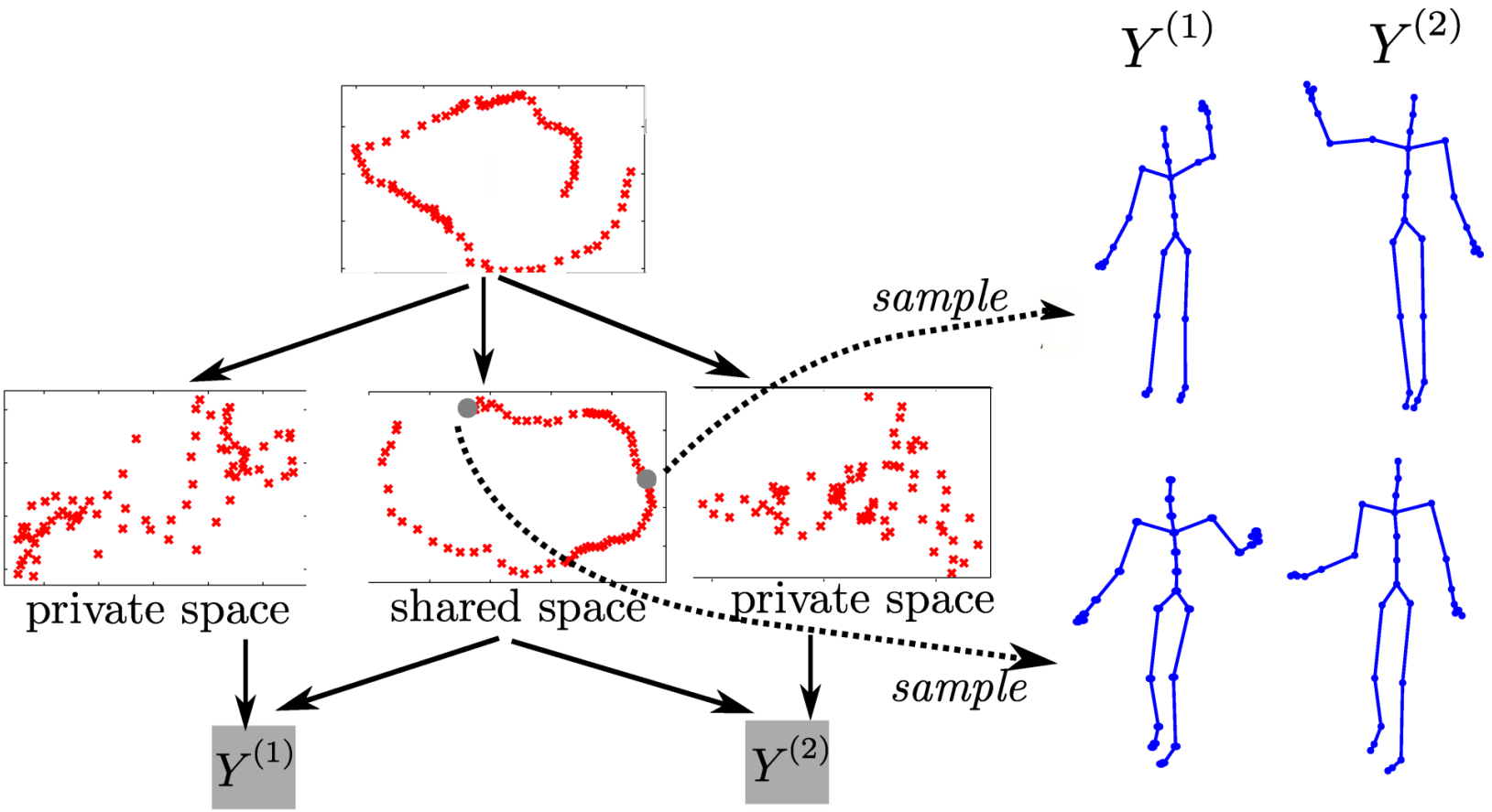


Face Demo

Motion Capture

- Revisit 'high five' data.
- This time allow model to learn structure, rather than imposing it.

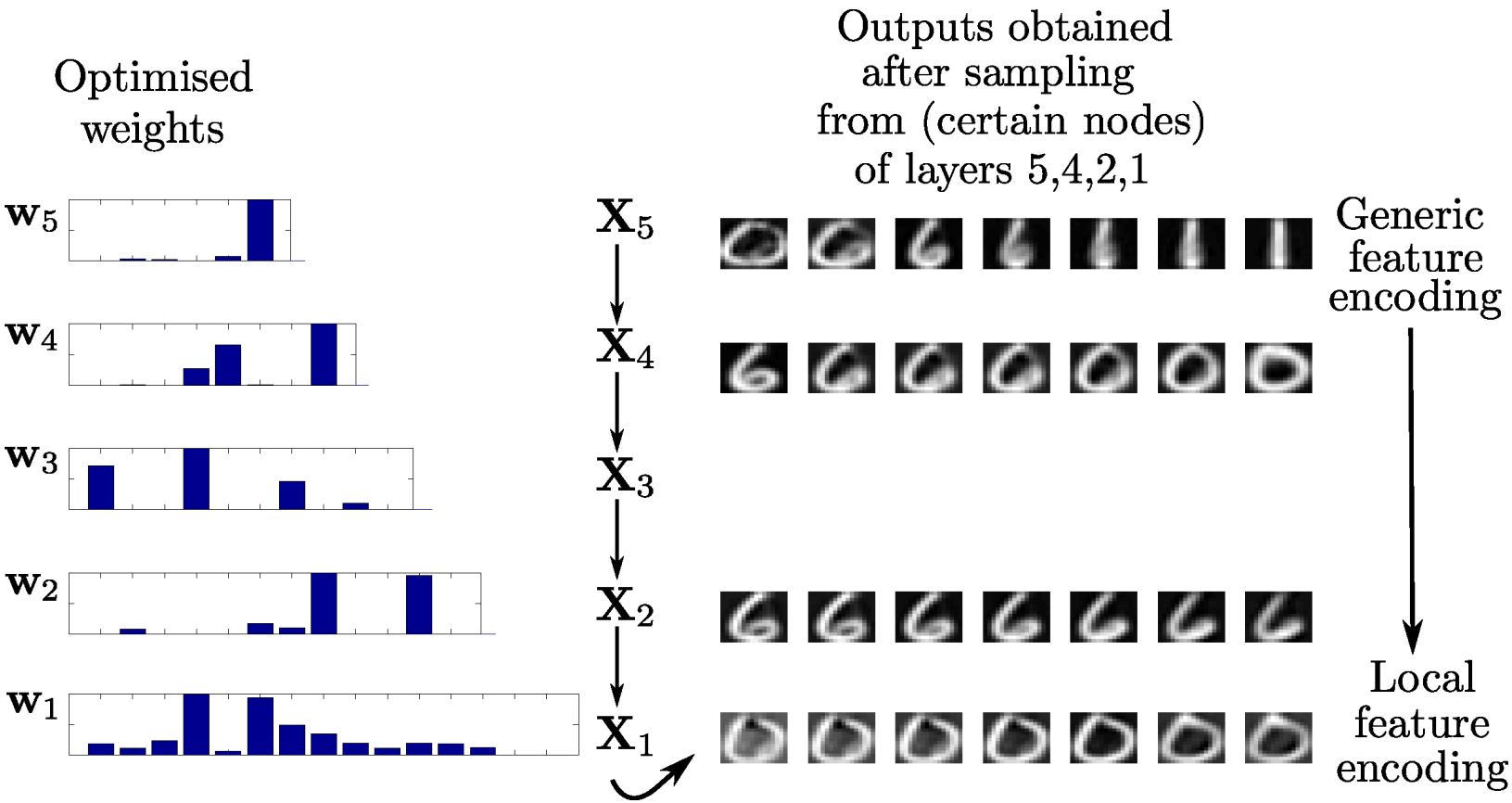
Deep hierarchies – motion capture



Digits Data Set

- Are deep hierarchies justified for small data sets?
- We can lower bound the evidence for different depths.
- For 150 6s, 0s and 1s from MNIST we found at least 5 layers are required.

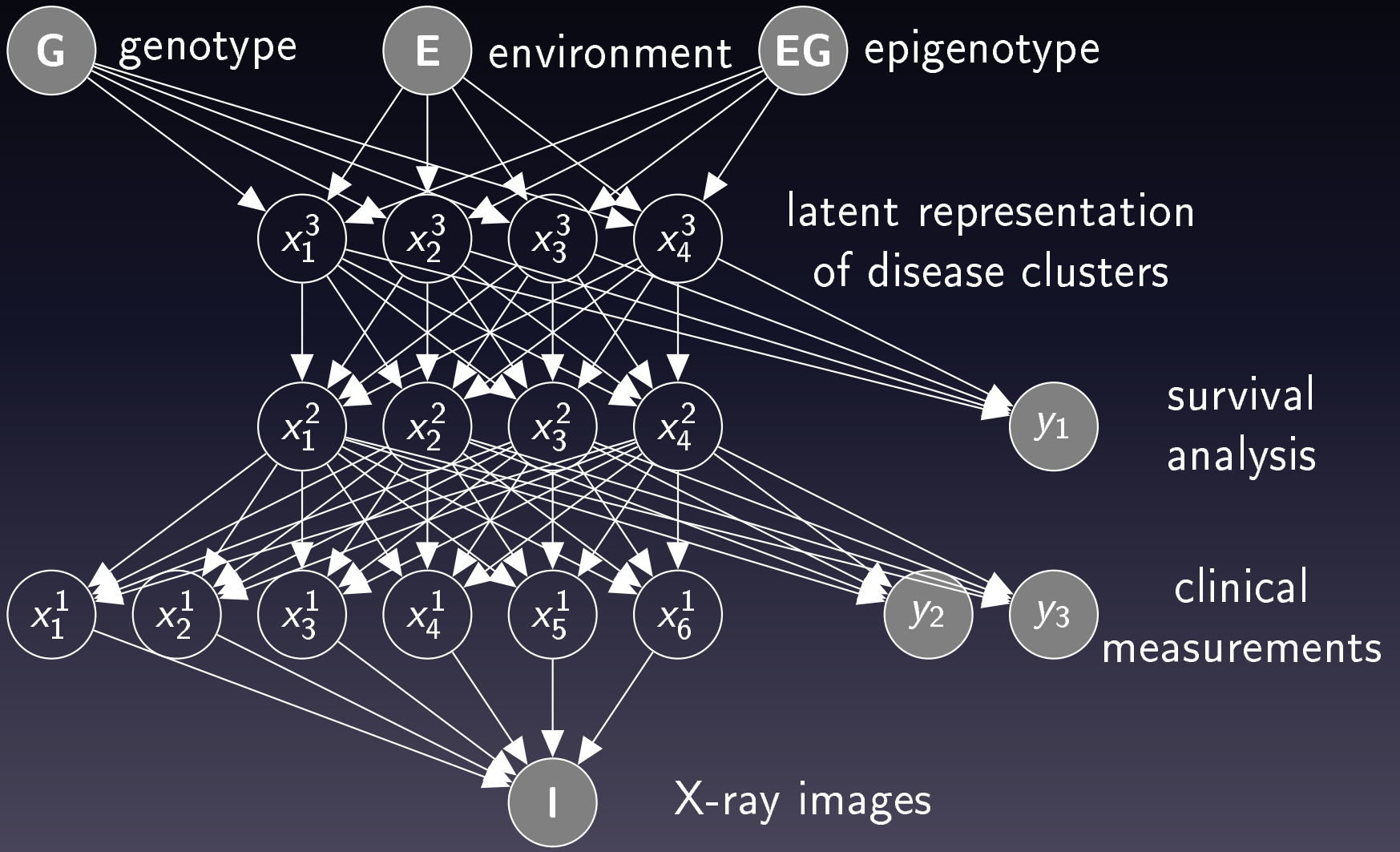
Deep hierarchies – MNIST



What Can We Do that Google Can't?

- Google's resources give them access to volumes of data (or Facebook, or Microsoft, or Amazon).
- Is there anything for Universities to contribute?
- Universities are the right place to deal with sensitive data for personalized health.
- These methodologies are part of that picture.

Deep Latent Variable Models for Personalized Health



Summary

- Deep models allow abstract representation of data sets at higher levels.
- Deep GPs allow structure learning.
- Current limitation is on data set size.
- Addressing this through work by James Hensman on Stochastic Variational Inference for GPs (NIPS Workshop Poster 'GPs for Big Data').
- Intention is to deploy these models for assimilating a wide range of data types in personalized health (text, survival times, images, genotype, phenotype).
- Requires population scale models with millions of features.

References I

Y. Bengio. Learning Deep Architectures for AI. *Found. Trends Mach. Learn.*, 2(1):1–127, Jan. 2009. ISSN 1935-8237. [DOI].

A. Damianou, C. H. Ek, M. K. Titsias, and N. D. Lawrence. Manifold relevance determination. In Langford and Pineau (2012). [PDF]. To appear.

A. Damianou and N. D. Lawrence. Deep Gaussian processes. In C. Carvalho and P. Ravikumar, editors, *Proceedings of the Sixteenth International Workshop on Artificial Intelligence and Statistics*, volume 31, AZ, USA, 2013. JMLR W&CP 31. [PDF].

C. H. Ek, J. Rihan, P. Torr, G. Rogez, and N. D. Lawrence. Ambiguity modeling in latent spaces. In A. Popescu-Belis and R. Stiefelhagen, editors, *Machine Learning for Multimodal Interaction (MLMI 2008)*, LNCS, pages 62–73. Springer-Verlag, 28–30 June 2008a. [PDF].

C. H. Ek, P. H. Torr, and N. D. Lawrence. Gaussian process latent variable models for human pose estimation. In A. Popescu-Belis, S. Renals, and H. Bourlard, editors, *Machine Learning for Multimodal Interaction (MLMI 2007)*, volume 4892 of *LNCS*, pages 132–143, Brno, Czech Republic, 2008b. Springer-Verlag. [PDF].

Z. Ghahramani, editor. *Proceedings of the International Conference in Machine Learning*, volume 24, 2007. Omnipress. [Google Books] .

G. E. Hinton. Products of experts. In *ICANN 99: Ninth international conference on artificial neural networks*, volume 1, pages 1–6. IEE Press, 1999.

G. E. Hinton and S. Osindero. A fast learning algorithm for deep belief nets. *Neural Computation*, 18: 2006, 2006.

A. Klami and S. Kaski. Local dependent components analysis. In Ghahramani (2007). [Google Books] .

A. Klami and S. Kaski. Probabilistic approach to detecting dependencies between data sets. *Neurocomputing*, 72:39–46, 2008.

J. Langford and J. Pineau, editors. *Proceedings of the International Conference in Machine Learning*, volume 29, San Francisco, CA, 2012. Morgan Kauffman. To appear.

N. D. Lawrence and A. J. Moore. Hierarchical Gaussian process latent variable models. In Ghahramani (2007), pages 481–488. [Google Books] . [PDF].

References II

G. Leen and C. Fyfe. A Gaussian process latent variable model formulation of canonical correlation analysis. Bruges (Belgium), 26-28 April 2006 2006.

D. J. C. MacKay. Introduction to Gaussian Processes. In C. M. Bishop, editor, *Neural Networks and Machine Learning*, volume 168 of *Series F: Computer and Systems Sciences*, pages 133–166. Springer-Verlag, Berlin, 1998.

R. Navaratnam, A. Fitzgibbon, and R. Cipolla. The joint manifold model for semi-supervised multi-valued regression. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE Computer Society Press, 2007.

R. Salakhutdinov and I. Murray. On the quantitative analysis of deep belief networks. In S. Roweis and A. McCallum, editors, *Proceedings of the International Conference in Machine Learning*, volume 25, pages 872–879. Omnipress, 2008.

A. P. Shon, K. Grochow, A. Hertzmann, and R. P. N. Rao. Learning shared latent structure for image synthesis and robotic imitation. In Y. Weiss, B. Schölkopf, and J. C. Platt, editors, *Advances in Neural Information Processing Systems*, volume 18, Cambridge, MA, 2006. MIT Press.

L. R. Tucker. An inter-battery method of factor analysis. *Psychometrika*, 23(2):111–136, 1958.

S. Virtanen, A. Klami, and S. Kaski. Bayesian CCA via group sparsity. In L. Getoor and T. Scheffer, editors, *Proceedings of the International Conference in Machine Learning*, volume 28, 2011.

A. G. Wilson, D. A. Knowles, and Z. Ghahramani. Gaussian process regression networks. In Langford and Pineau (2012). To appear.