Deep Learning: Looking Forward

Yoshua Bengio
U. Montreal

July 16th, 2013

Microsoft Research Faculty Summit 2013, Bellevue, WA

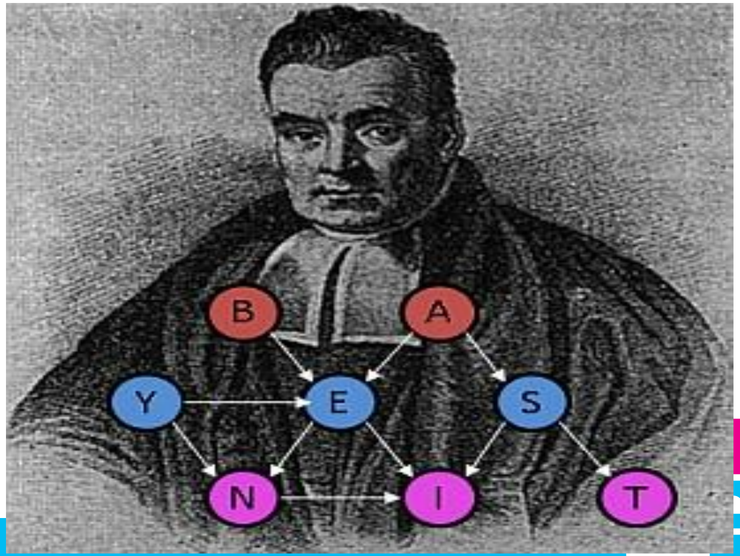
Université de Montréal

LISA

Representation Learning

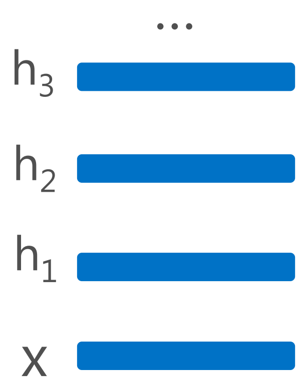
- Good input features essential for successful ML *(feature engineering = 90% of effort in industrial ML)*
- Handcrafting features vs learning them

- Representation learning: guesses
  the features / factors / causes =
  good representation.



Deep Representation Learning

Learn multiple levels of representation of increasing complexity/abstraction

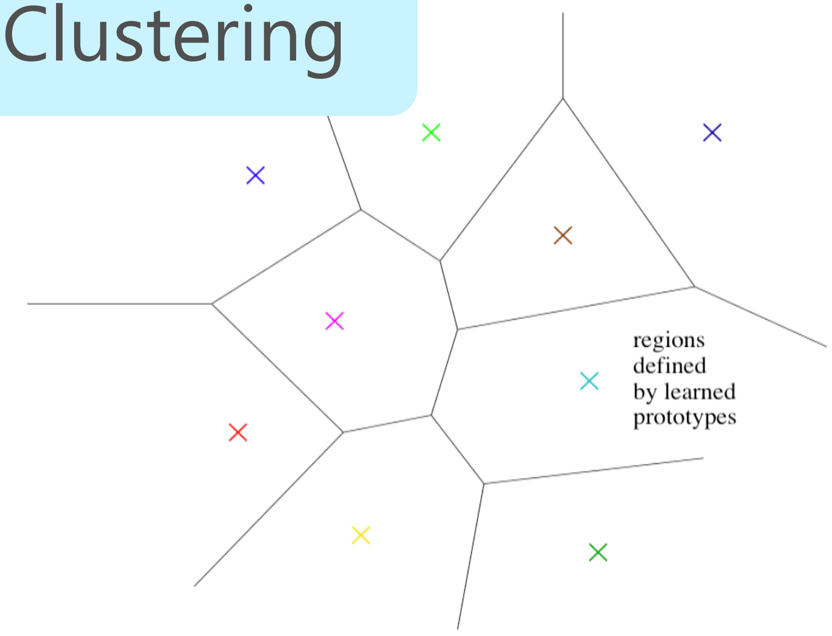


- potentially exponential gain in expressive power
- brains are deep
- humans organize knowledge in a compositional way

- Better mixing in the space of deeper representations (Bengio et al, ICML 2013)
- **They work! SOTA on industrial-scale AI tasks (object recognition, speech recognition, language modeling, music modeling)**

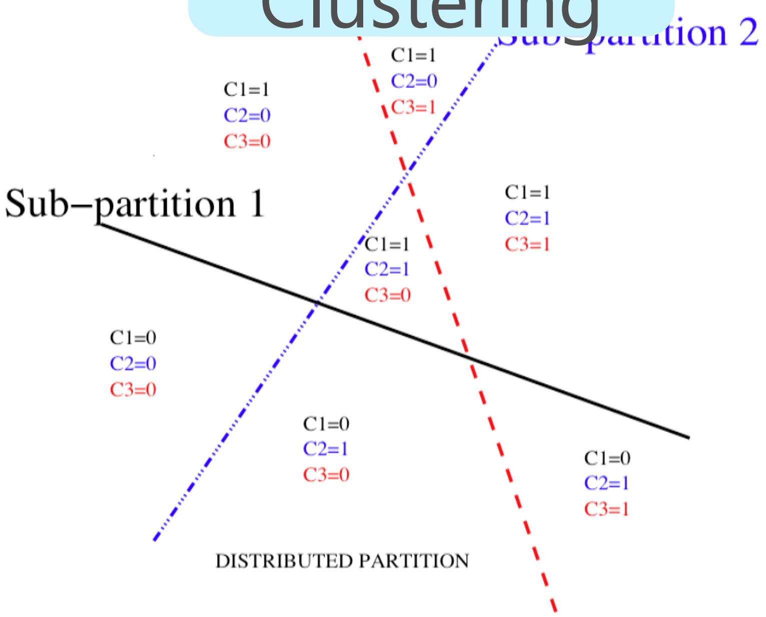
The need for distributed representations

Clustering



LOCAL PARTITION

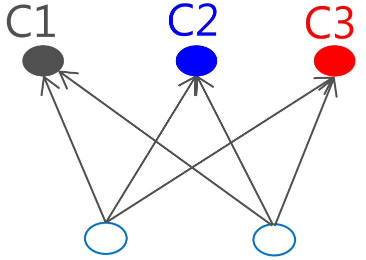
Multi-Clustering



Each parameter influences many regions, not just local neighbors

distinguishable regions grows almost exponentially with # parameters

Learning a set of features that are not mutually exclusive can be exponentially more statistically efficient than nearest-neighbor-like or clustering-like models



Learning multiple levels of representation

(Lee, Largman, Pham & Ng, NIPS 2009)
(Lee, Grosse, Ranganath & Ng, ICML 2009)

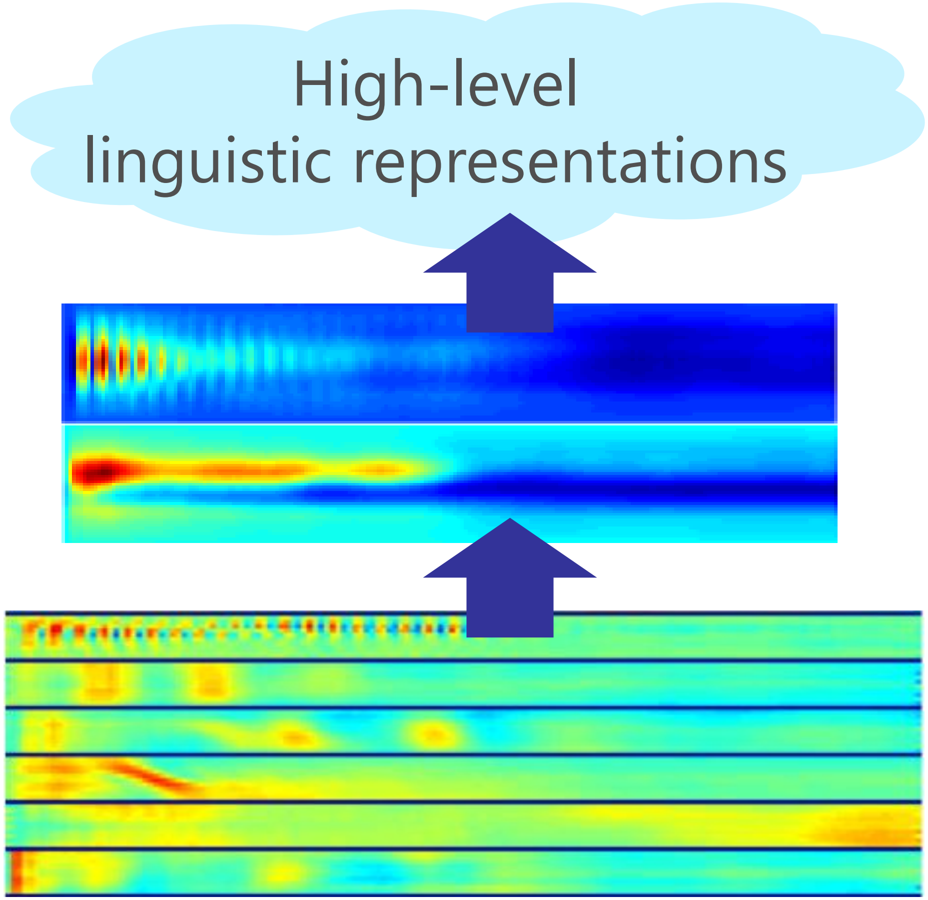
Successive model layers learn deeper intermediate representations

Layer 3

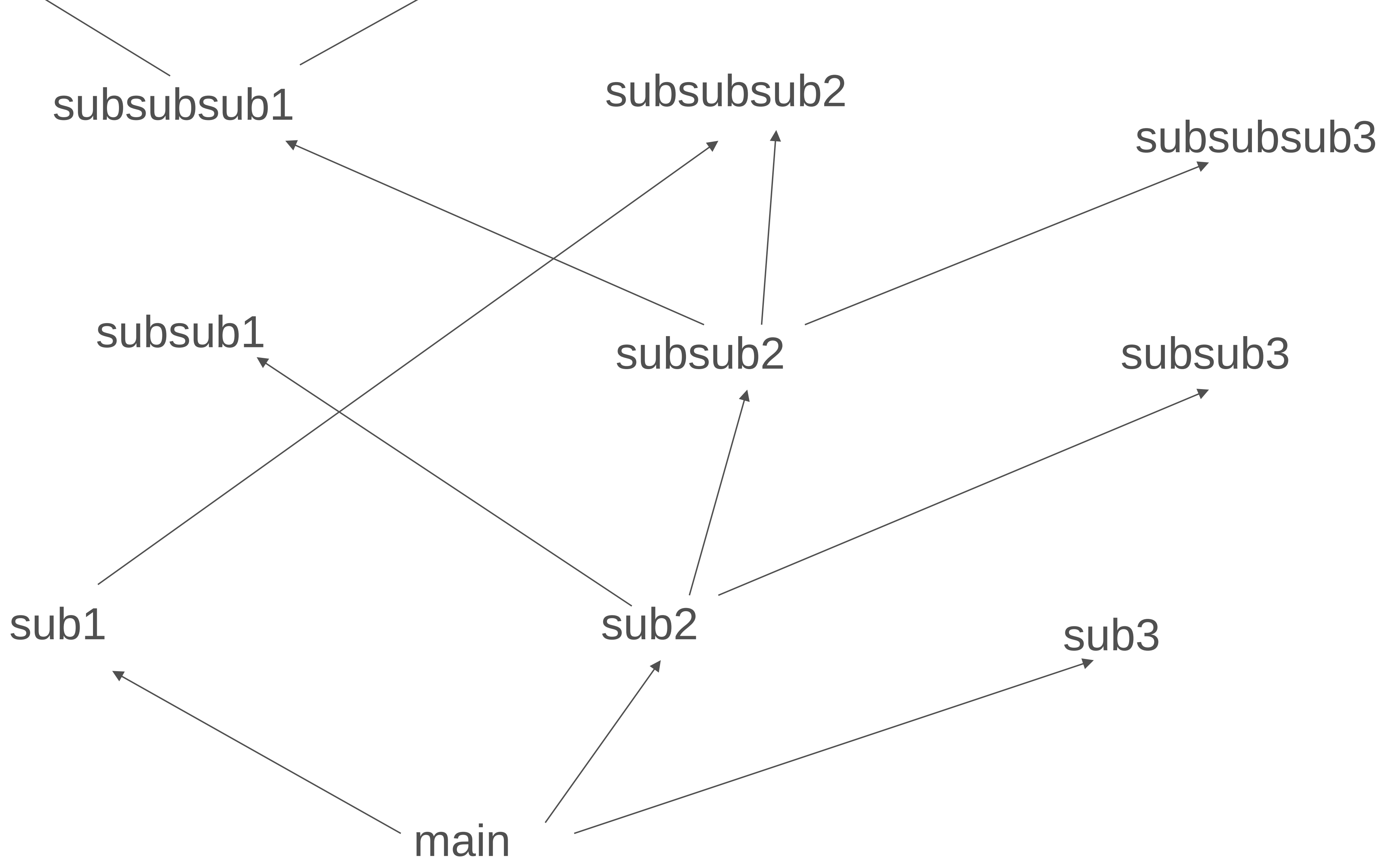
Parts combine
to form objects

Layer 2

Layer 1



Prior: underlying factors & concepts compactly expressed w/ multiple levels of abstraction



“Deep” computer program

subroutine1 includes subsub1 code and subsub2 code and subsubsub1 code

subroutine2 includes subsub2 code and subsub3 code and subsubsub3 code and …

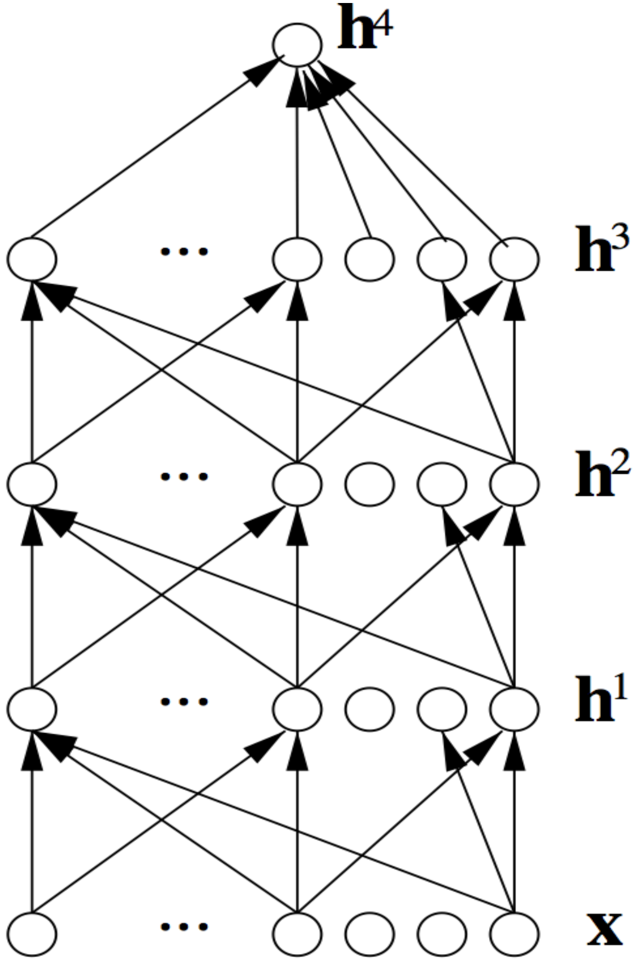
main

“Shallow” computer program

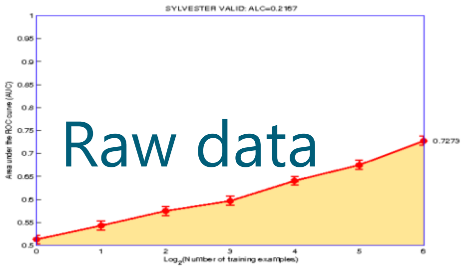
Deep Supervised Neural Nets

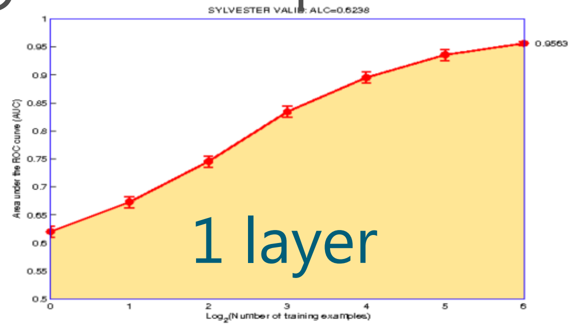
- We can now train them even without unsupervised pre-training, thanks to better initialization and non-linearities (rectifiers, maxout) and they can generalize well with large labeled sets and dropout.

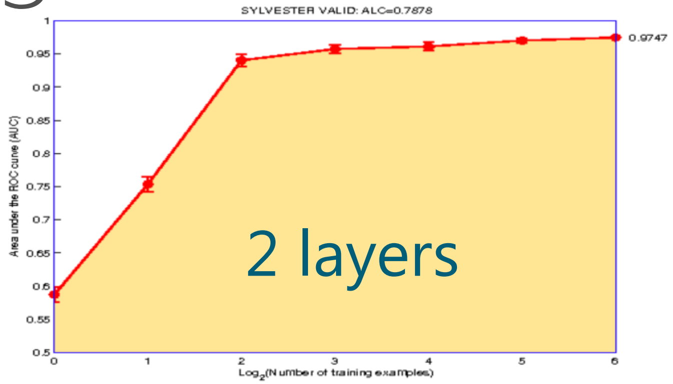
- Unsupervised pre-training still useful for rare classes, transfer, smaller labeled sets, or as an extra regularizer.



Unsupervised and Transfer Learning Challenge + Transfer Learning Challenge: Deep Learning 1st Place

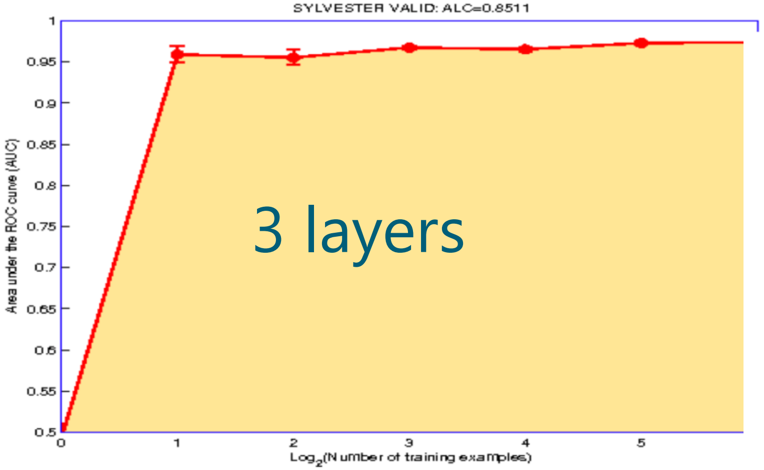


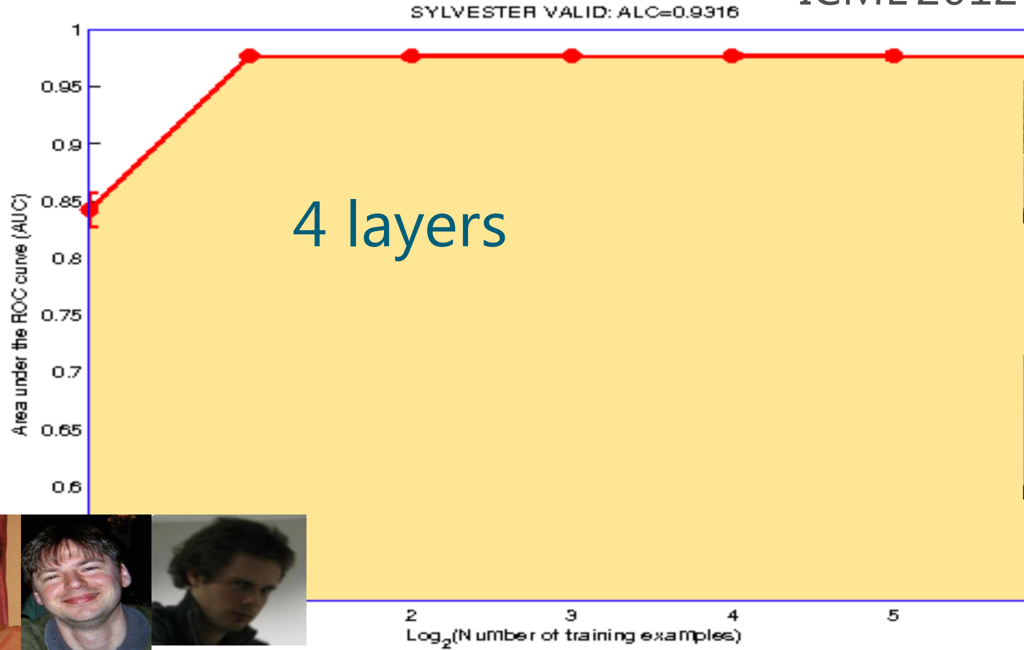


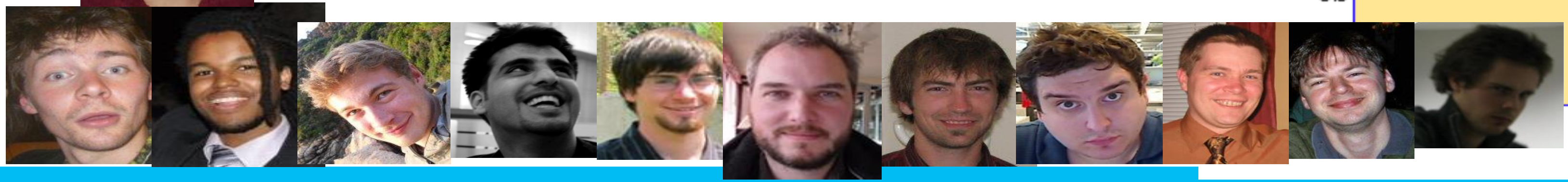


NIPS'2011 Transfer Learning Challenge Paper: ICML'2012

ICML'2011 workshop on Unsup. & Transfer Learning

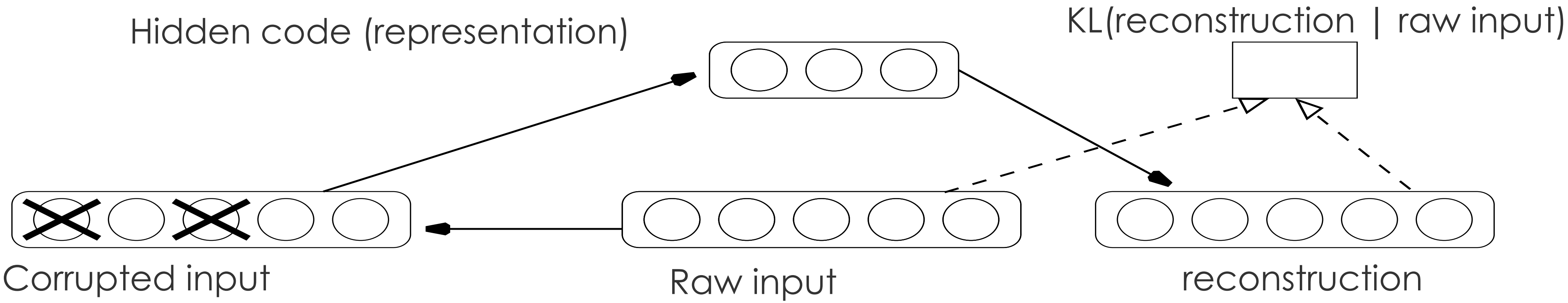






Denoising Auto-Encoder

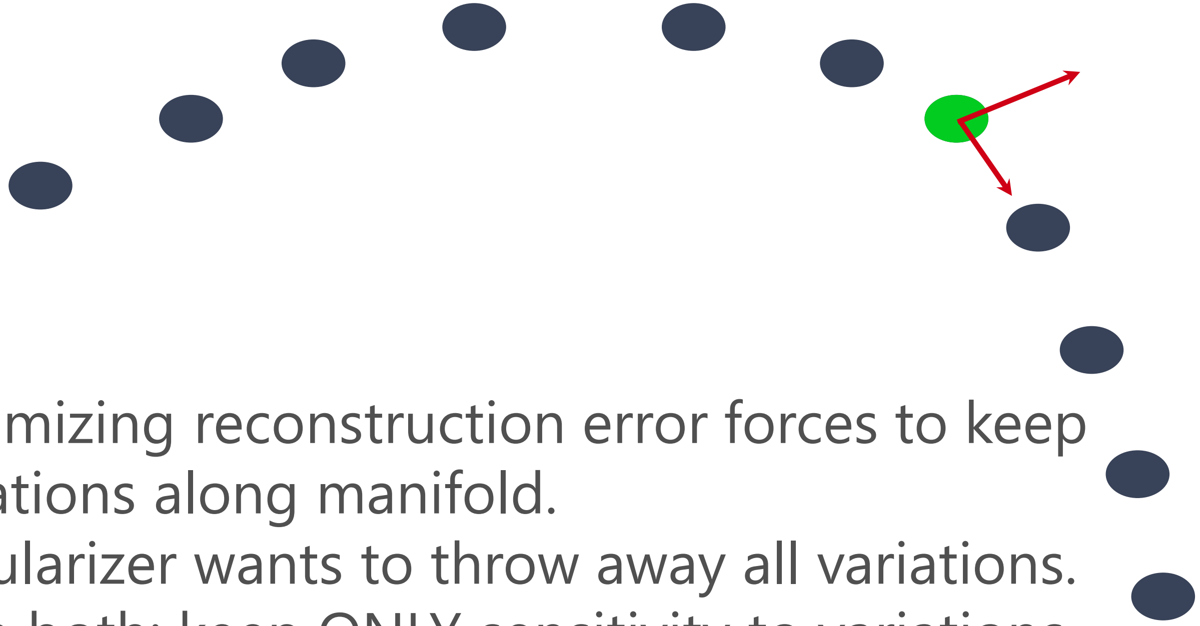
(Vincent et al 2008)



- Corrupt the input
- Try to reconstruct the uncorrupted input

- Models the input density through a form of score matching (Vincent 2011, Alain & Bengio ICLR 2013) or as the transition kernel of a Markov chain (Bengio et al, arxiv 2013 "Generalized Denoising Auto-Encoders as Generative Models")

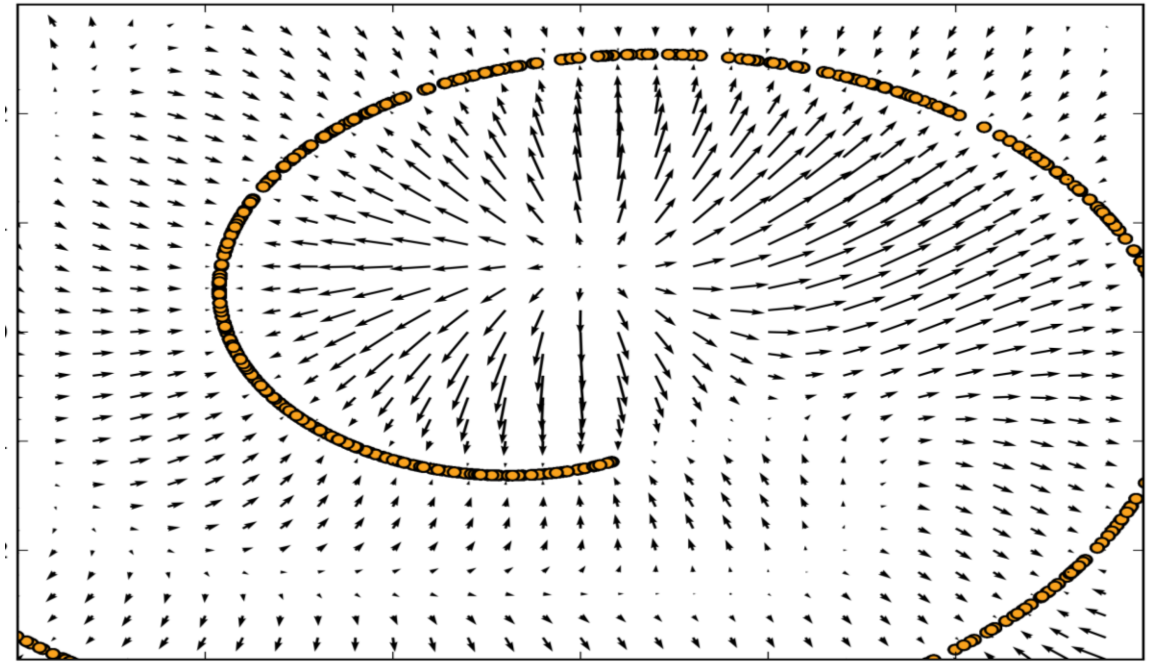
Regularized Auto-Encoders Learn Salient Variations, like non-linear PCA with shared parameters

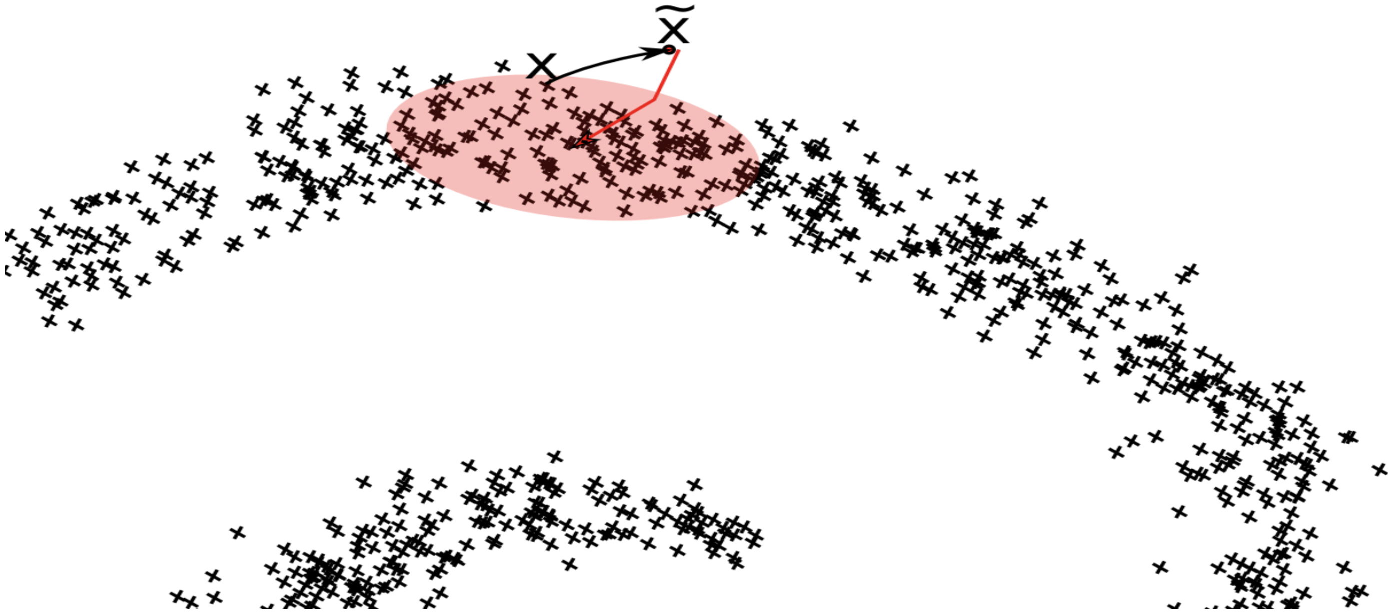


- Minimizing reconstruction error forces to keep variations along manifold.
- Regularizer wants to throw away all variations.
- With both: keep ONLY sensitivity to variations ON the manifold.

Regularized Auto-Encoders Learn a Vector Field or a Markov Chain Transition Distribution

- (Bengio, Vincent & Courville, TPAMI 2013) review paper
- (Alain & Bengio ICLR 2013; Bengio et al, arxiv 2013)



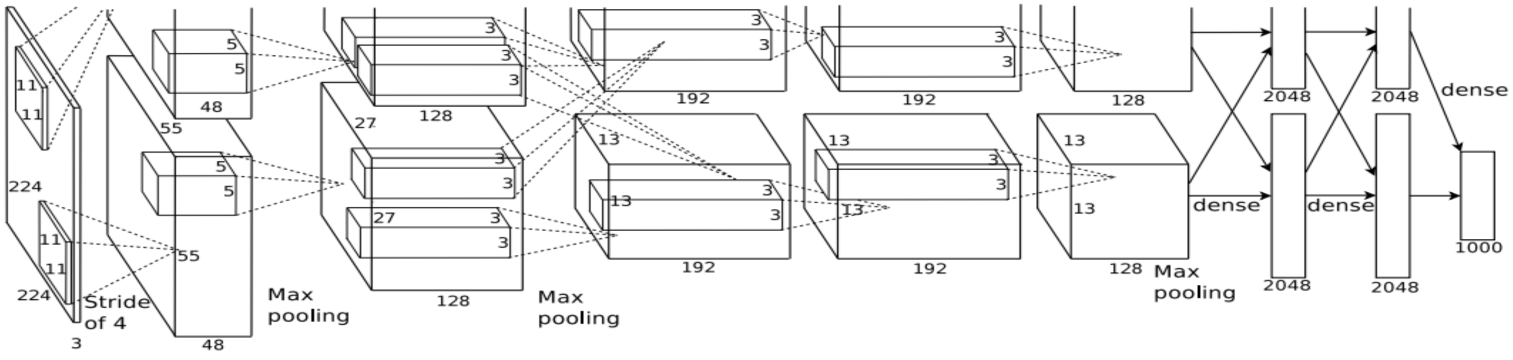


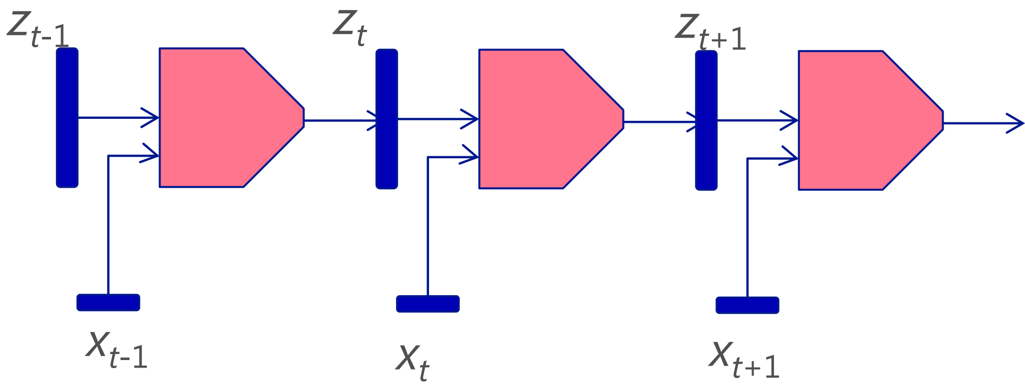
Stochastic Neurons as Regularizer: Improving neural networks by preventing co-adaptation of feature detectors (Hinton et al 2012, arXiv)

- Dropouts trick: during training multiply neuron output by random bit (p=0.5), during test by 0.5
- Used in deep supervised networks
- Similar to denoising auto-encoder, but corrupting every layer
- Works better with rectifiers, even better with maxout (Goodfellow et al. ICML 2013)
- Equivalent to averaging over exponentially many architectures
  - Used by Krizhevsky et al to break through ImageNet SOTA
  - Also improves SOTA on CIFAR-10 (18→16% err)
  - Knowledge-free MNIST with DBMs (.95→.79% err)
  - TIMIT phoneme classification (22.7→19.7% err)

Temporal & Spatial Inputs: Convolutional & Recurrent Nets

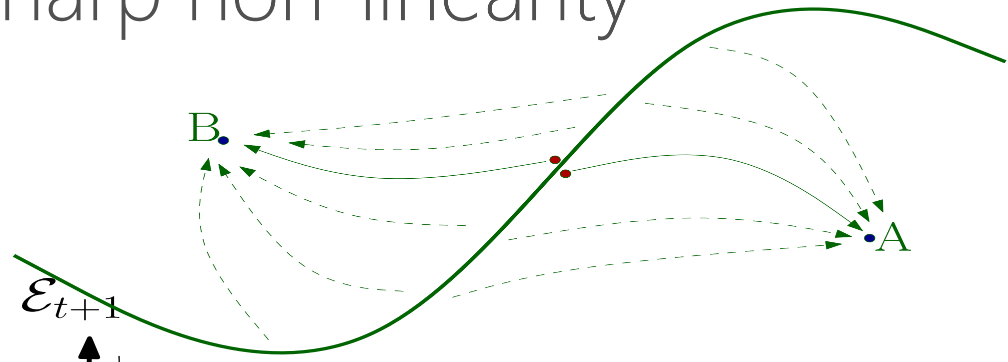
- Local connectivity across time/space
- Sharing weights across time/space (translation equivariance)
- Pooling (translation invariance, cross-channel pooling for others)
- Recurrent nets can summarize information from the past
- Bidirectional recurrent nets can also summarize information from the future

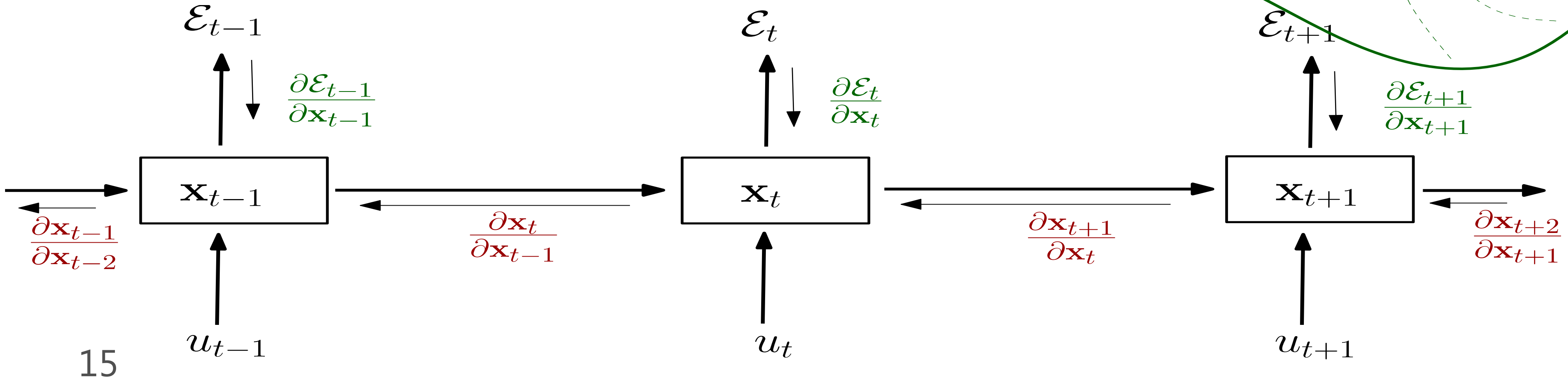




The Optimization Challenge in Deep / Recurrent Nets

- Higher-level abstractions require highly non-linear transformations to be learned
- Sharp non-linearities are difficult to learn by gradient
- Composition of many non-linearities = sharp non-linearity
- Exploding or vanishing gradients



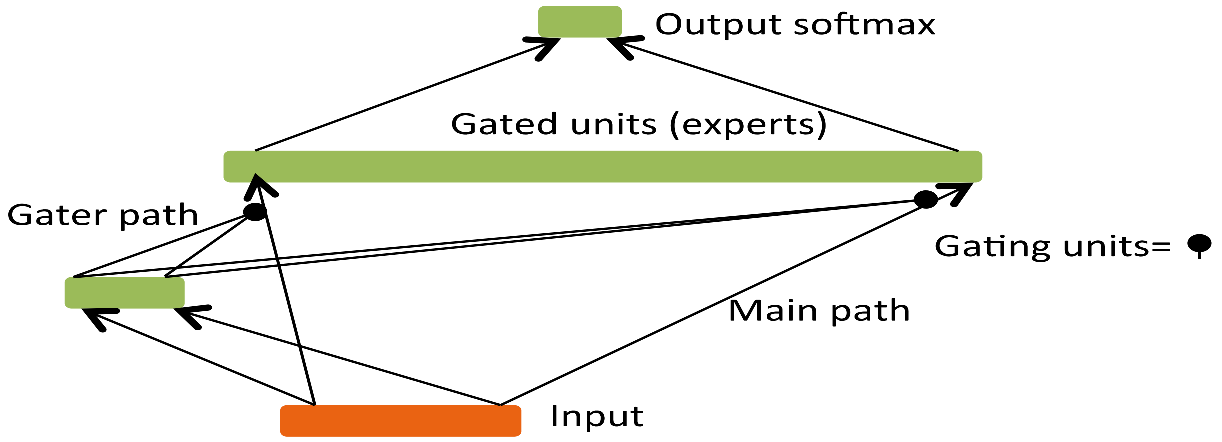


Deep Learning Challenges (Bengio, arxiv 1305.0445 Deep learning of representations: looking forward)

- Computational Scaling
- Optimization & Underfitting
- Approximate Inference & Sampling
- Disentangling Factors of Variation
- Reasoning & One-Shot Learning of Facts

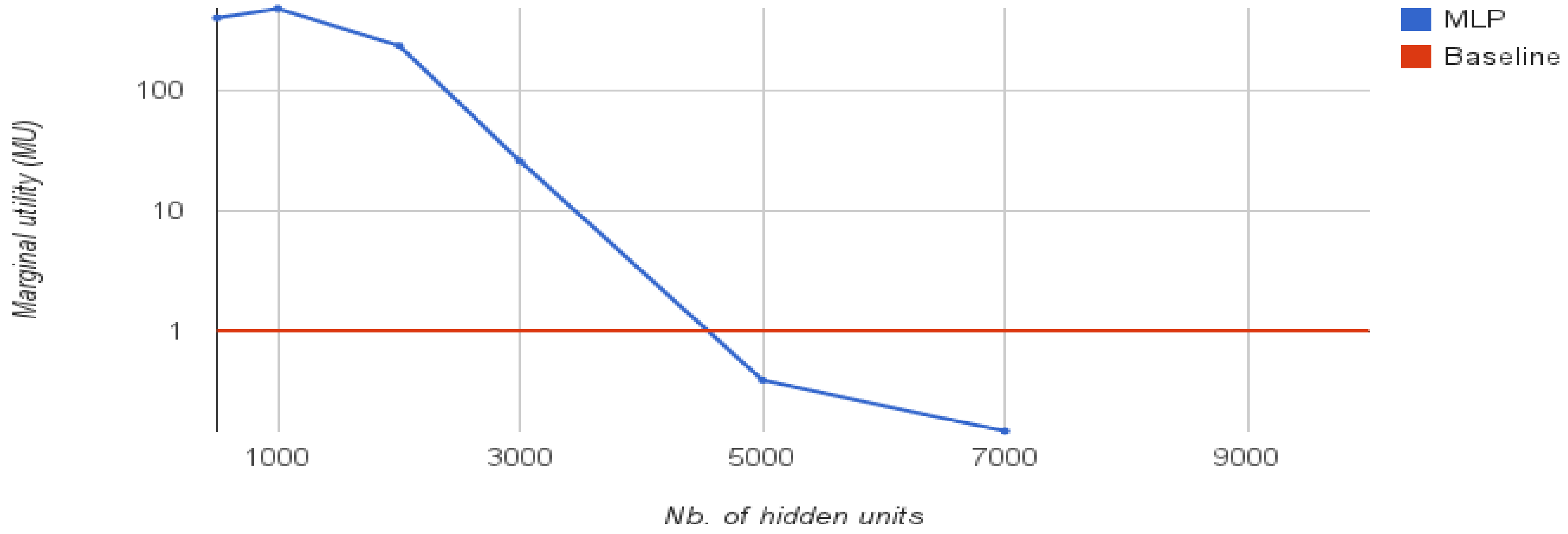
Conditional Computation

- Deep nets vs decision trees
- Hard mixtures of experts
- Conditional computation for deep nets: sparse distributed gaters selecting combinatorial subsets of a deep net
- Challenges:
  - Back-prop through hard decisions
  - Gated architectures exploration
- Symmetry breaking to reduce ill-conditioning



Optimization & Underfitting

- On large datasets, major obstacle is underfitting
- Marginal utility of wider MLPs decreases quickly below memorization baseline



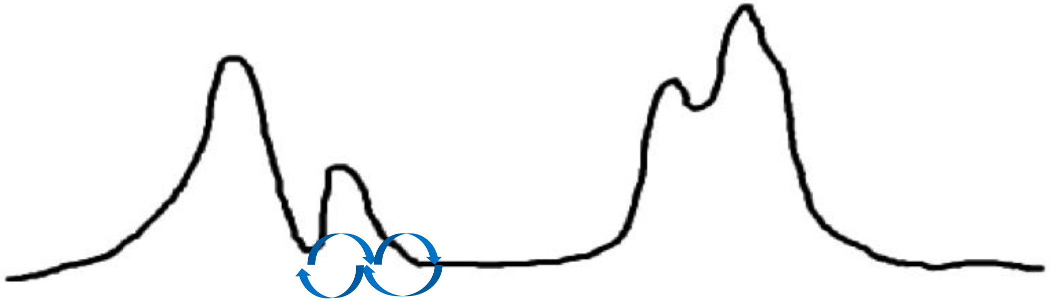
- Current limitations: local minima or ill-conditioning?
- Adaptive learning rates and stochastic 2nd order methods
- Conditional comp. & sparse gradients → better conditioning: when some gradients are 0, many cross-derivatives are also 0.

Inference & Sampling

- Currently for unsupervised learning & structured output models
- P(h|x) intractable because of many important modes
- MAP, Variational, MCMC approximations limited to 1 or few modes

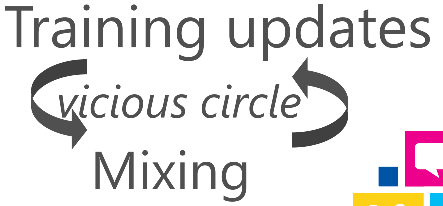
- Approximate inference can hurt learning

  (Kulesza & Pereira NIPS'2007)



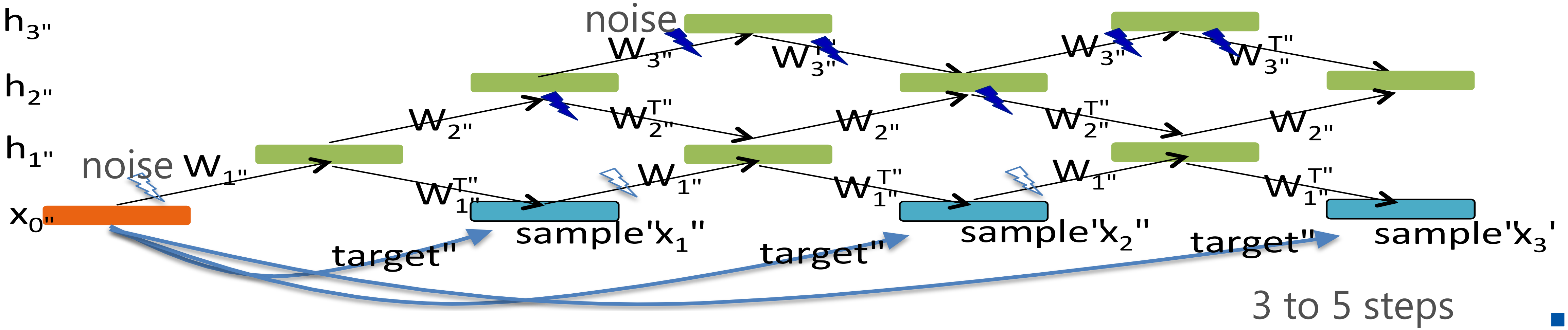
- Mode mixing harder as training progresses

  (Bengio et al ICML 2013)

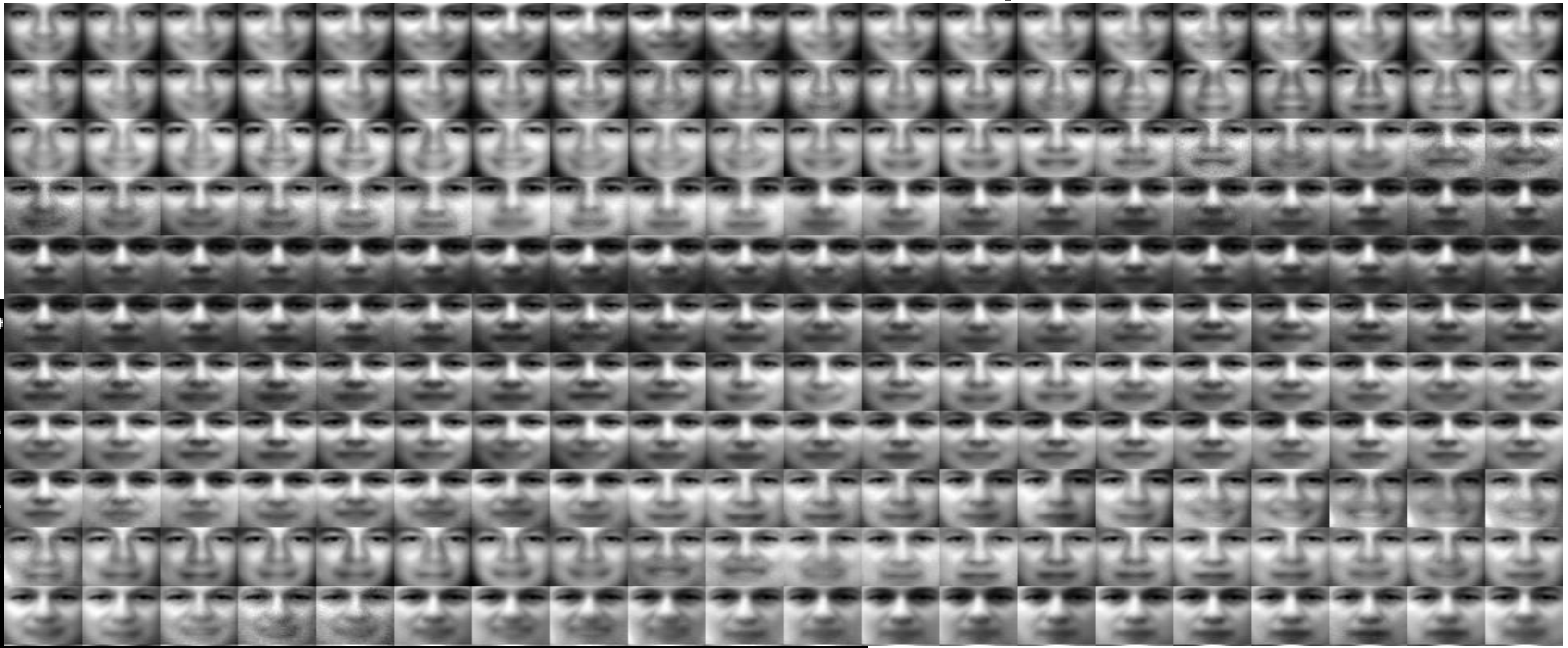


Learning Computational Graphs

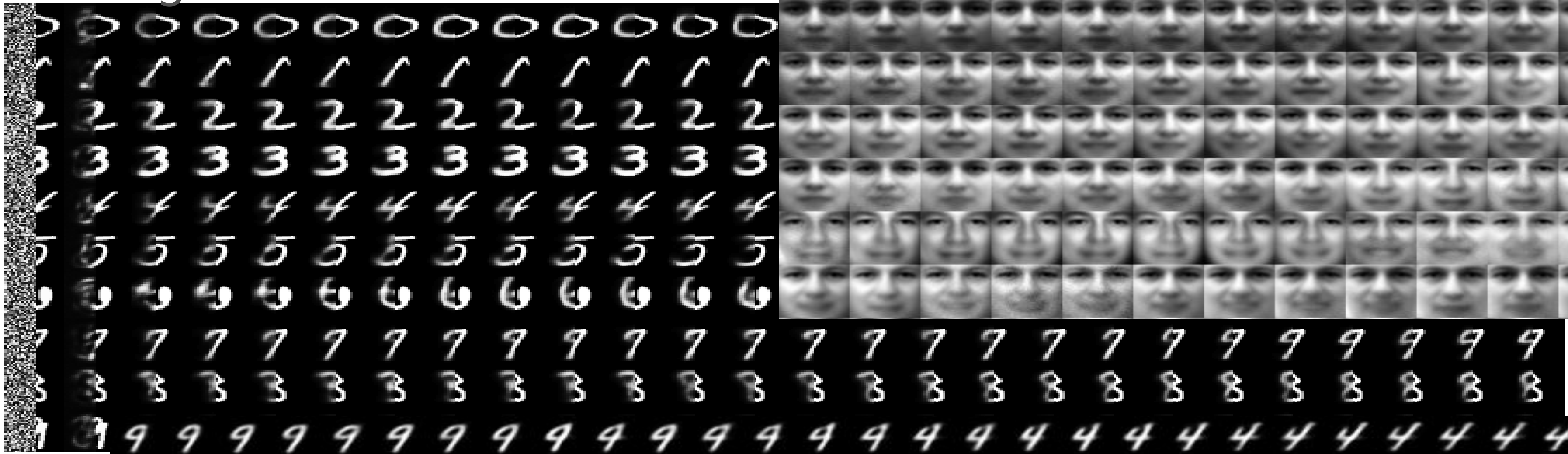
- Deep Stochastic Generative Networks (GSNs) trainable by backprop (Bengio & Laufer, arxiv 1306.1091)
- Avoid any explicit latent variables whose marginalization is intractable, instead train a stochastic computational graph that generates the right {conditional} distribution.

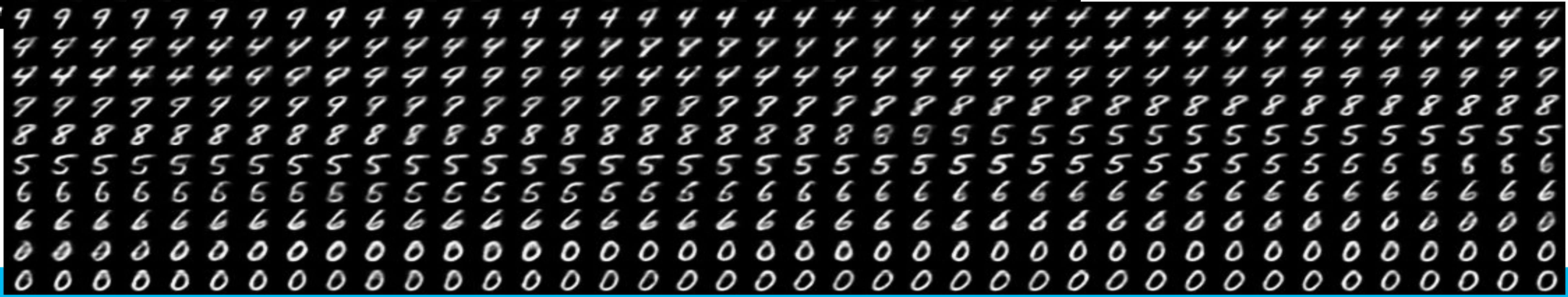


GSN Experiments: Consecutive Samples



Filling-in the LHS





Conclusions

- Deep Learning & Representation Learning have matured
  - Int. Conf. on Learning Representation 2013 a huge success!
- Industrial strength applications in place (Google, Microsoft)

- Room for improvement:
  - Scaling computation even more
  - Better optimization
  - Getting rid of intractable inference (in the works!)
  - Coaxing the models into more disentangled abstractions
  - Learning to reason from incrementally added facts

Merci! Questions?

LISA team: