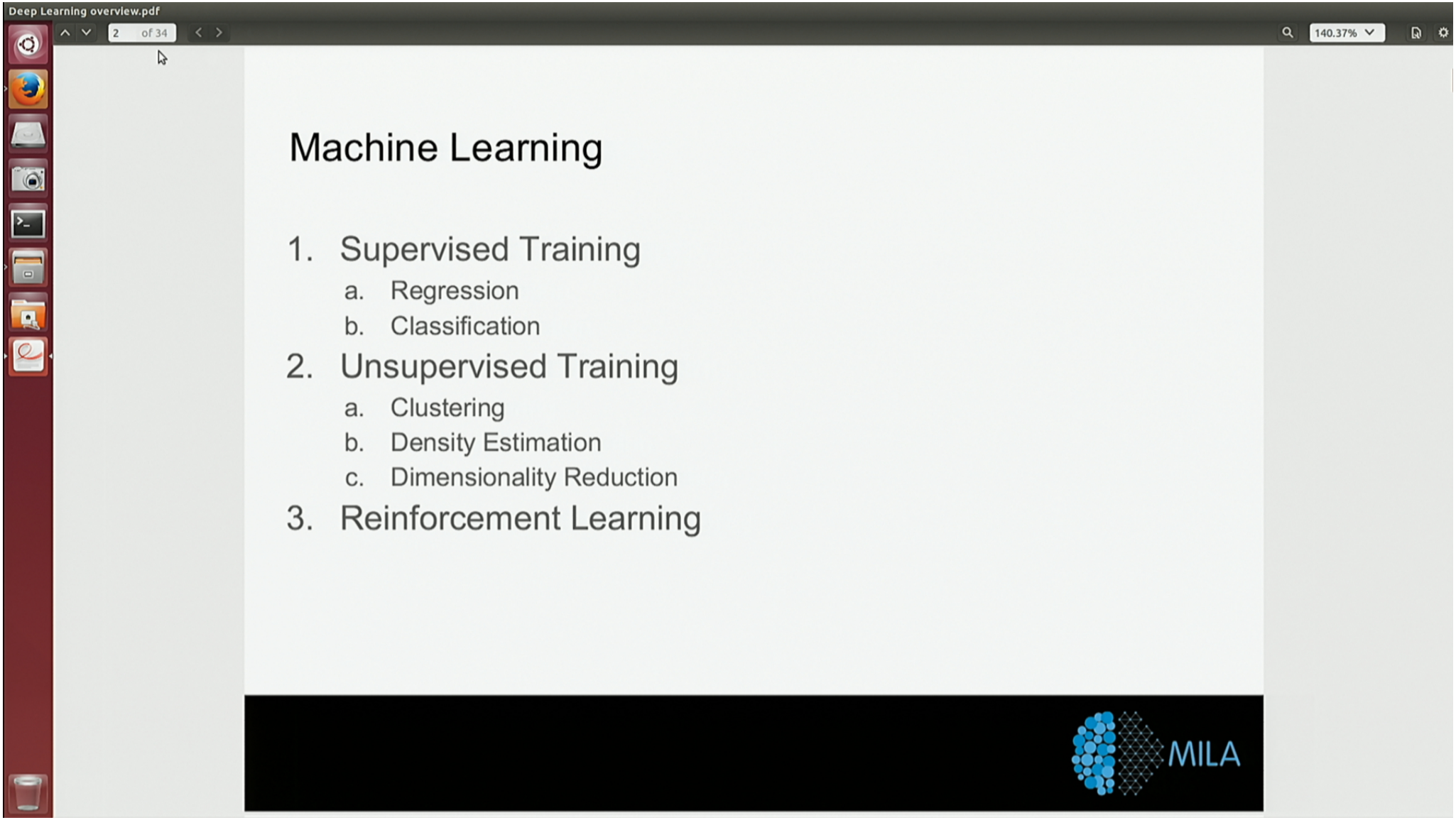
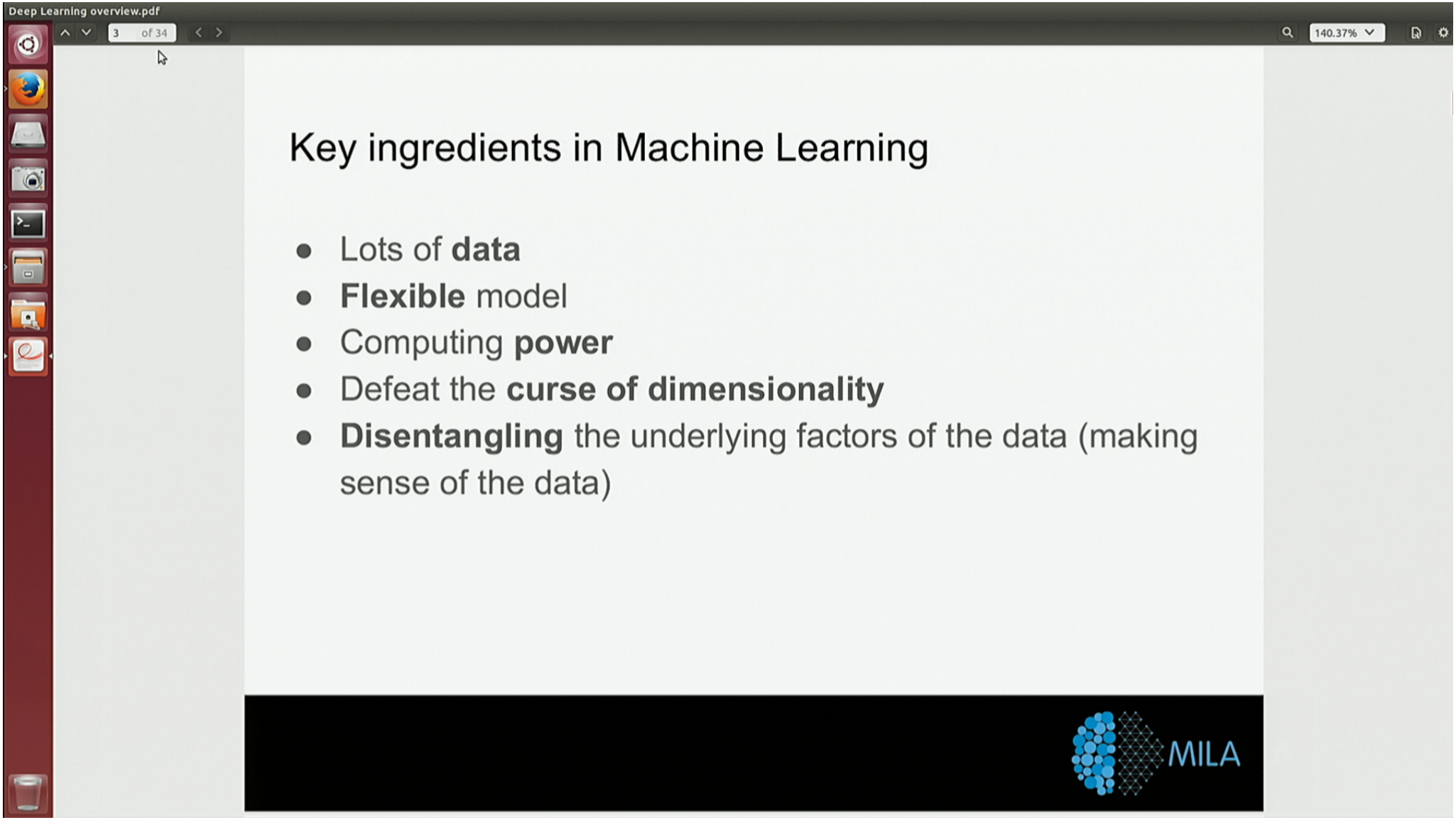
Title: Deep Learning: An Overview

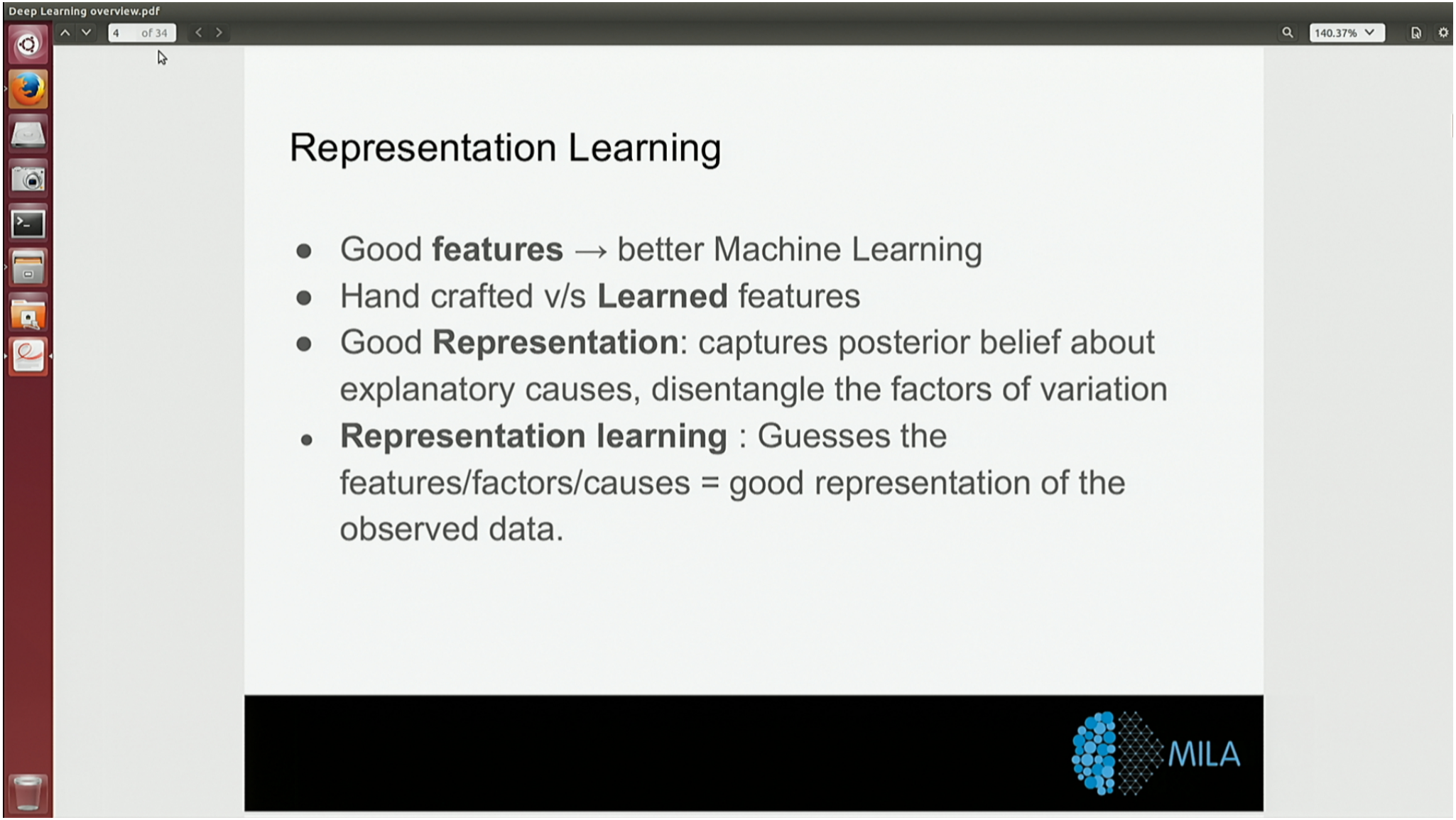
Date: Aug 09, 2016 03:00 PM

URL: http://pirsa.org/16080011

Abstract:

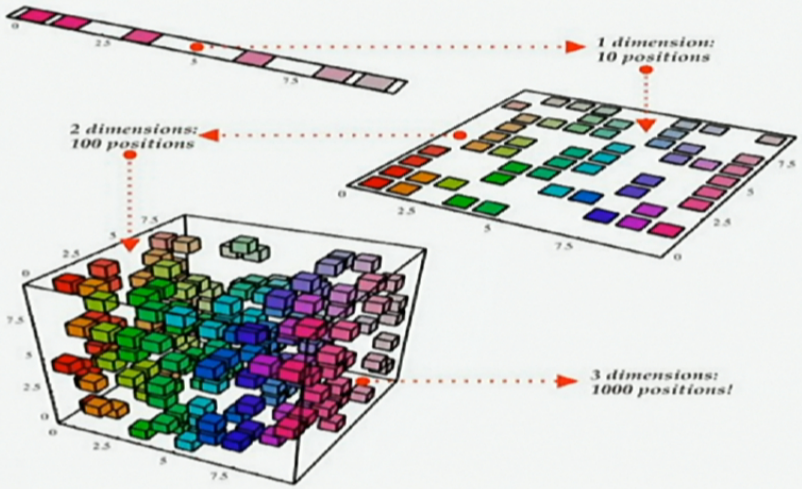

Deep Learning overview.pdf
2 of 34
140.37%
Machine Learning
1. Supervised Training
a. Regression
b. Classification
2. Unsupervised Training
a. Clustering
b. Density Estimation
c. Dimensionality Reduction
3. Reinforcement Learning
MILA


Deep Learning overview.pdf
3 of 34
140.37%
Key ingredients in Machine Learning
Lots of data
Flexible model
Computing power
Defeat the curse of dimensionality
Disentangling the underlying factors of the data (making sense of the data)
MILA

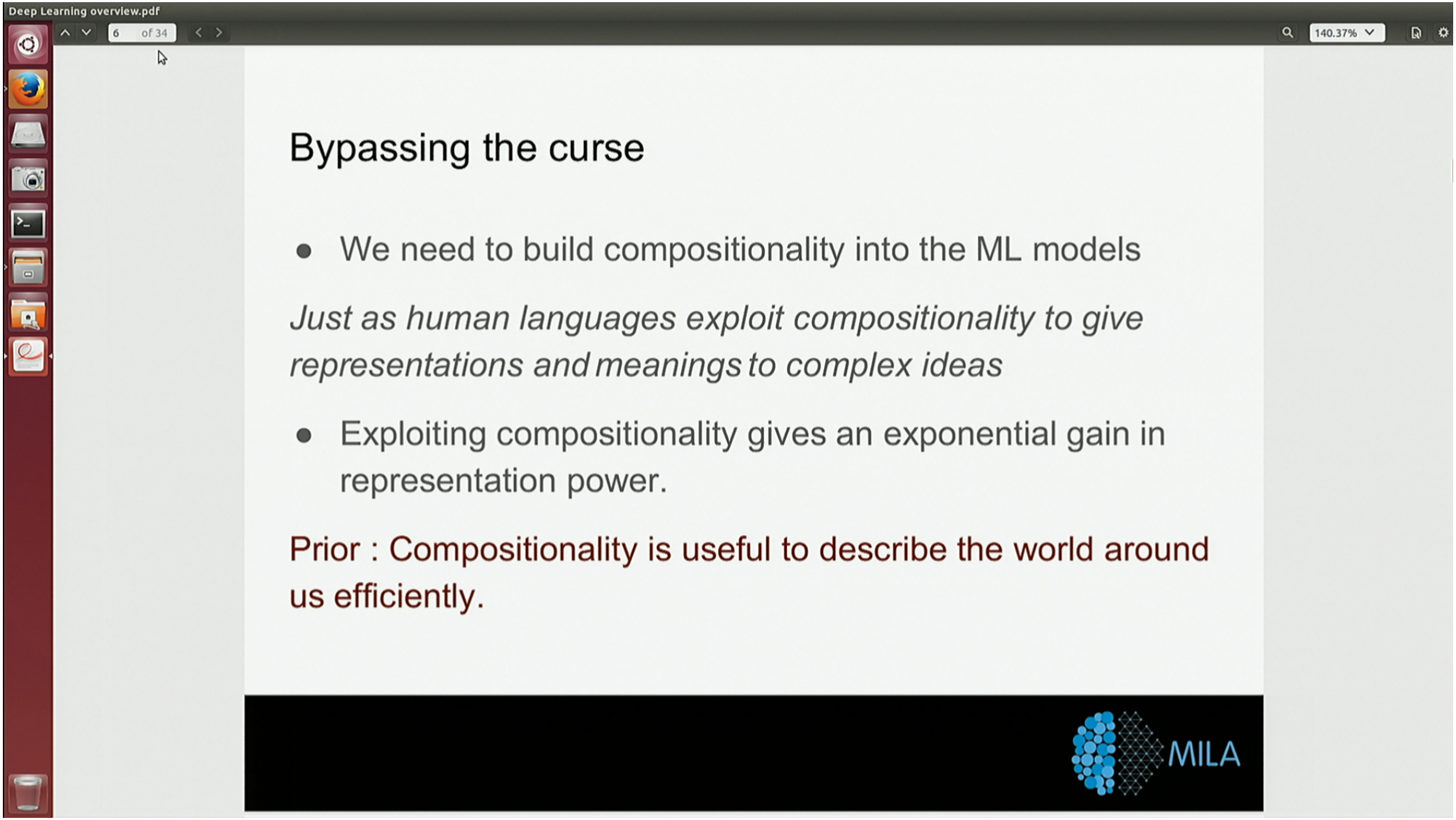

Deep Learning overview.pdf
4 of 34
140.37%
Representation Learning
Good features → better Machine Learning
Hand crafted v/s Learned features
Good Representation: captures posterior belief about explanatory causes, disentangle the factors of variation
Representation learning : Guesses the features/factors/causes = good representation of the observed data.
MILA

How to compose the hidden features
Fighting against: Curse of dimensionality

- To generalize locally: need representative example of all relevant variations
- **Classical approach: Hope for a smooth enough target function, or make it smooth by handcrafting good features.**



MILA


Deep Learning overview.pdf
6 of 34
140.37%
Bypassing the curse
We need to build compositionality into the ML models
Just as human languages exploit compositionality to give representations and meanings to complex ideas
Exploiting compositionality gives an exponential gain in representation power.
Prior : Compositionality is useful to describe the world around us efficiently.
MILA

Non-distributed representations

- Clustering, Nearest-Neighbors, decision trees, etc.
- Parameters for each distinguishable region.
- # of distinguishable region is linear in terms of parameters.

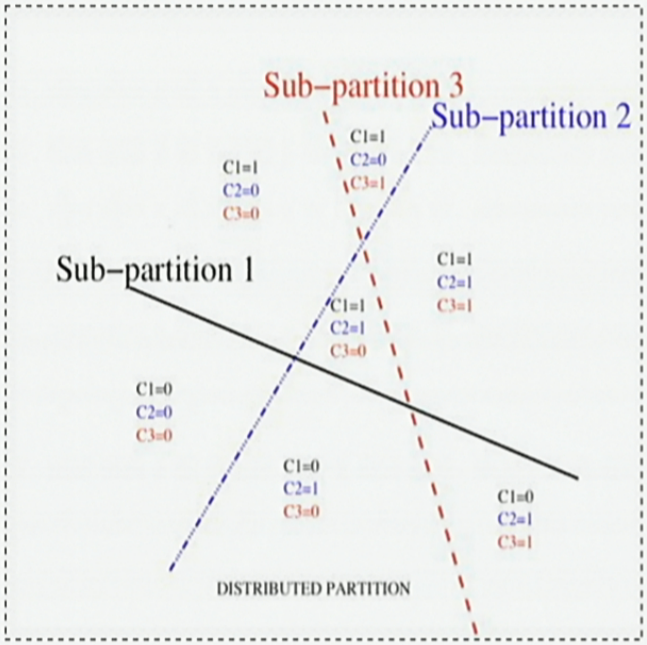
No non trivial generalization to regions without examples!!

MILA

The need for distributed representations

- PCA, RBM's, Neural Networks, Deep Learning etc.
- Each parameter influences many regions not just local regions.
- Learning a set of features that are not mutually exclusive can be exponentially more statistically efficient than having clustering like models.

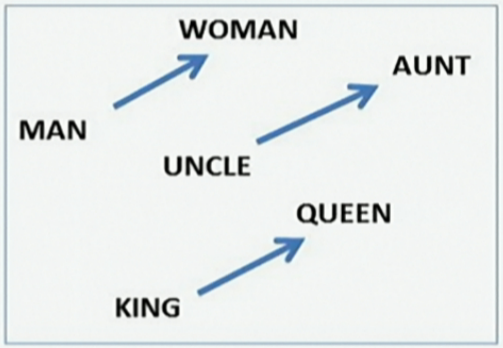
GENERALIZE NON LOCALLY TO UNSEEN EXAMPLES!!!!!!!!!

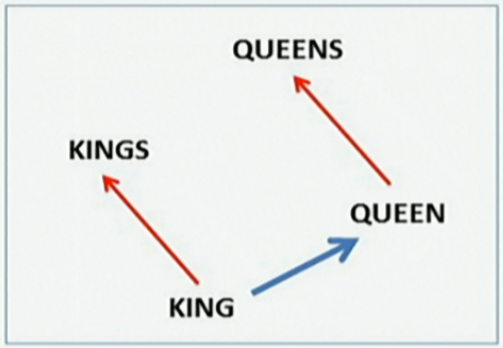


MILA

Analogical Representations for Free

- Semantic relations appear as linear relationships in the space of learned representations.
- King – Queen ≈ Man – Woman
- Paris - France + Italy = Rome





(Mikolov et al., NAACL HLT, 2013)

MILA

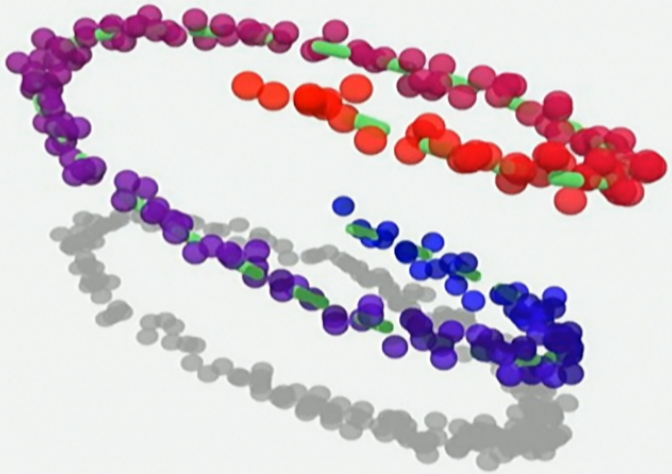
Deep Learning overview.pdf

10 of 34

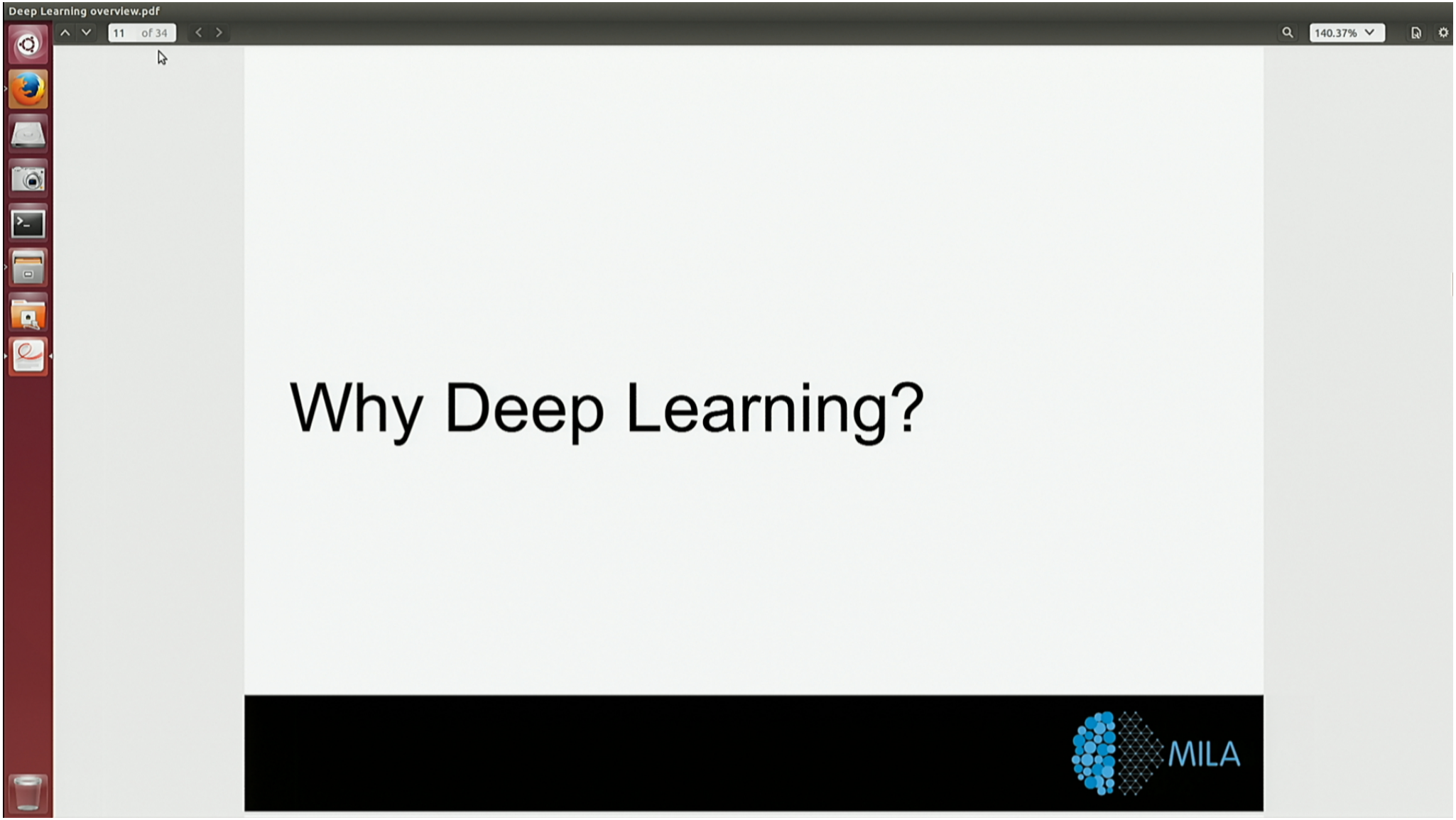
140.37%

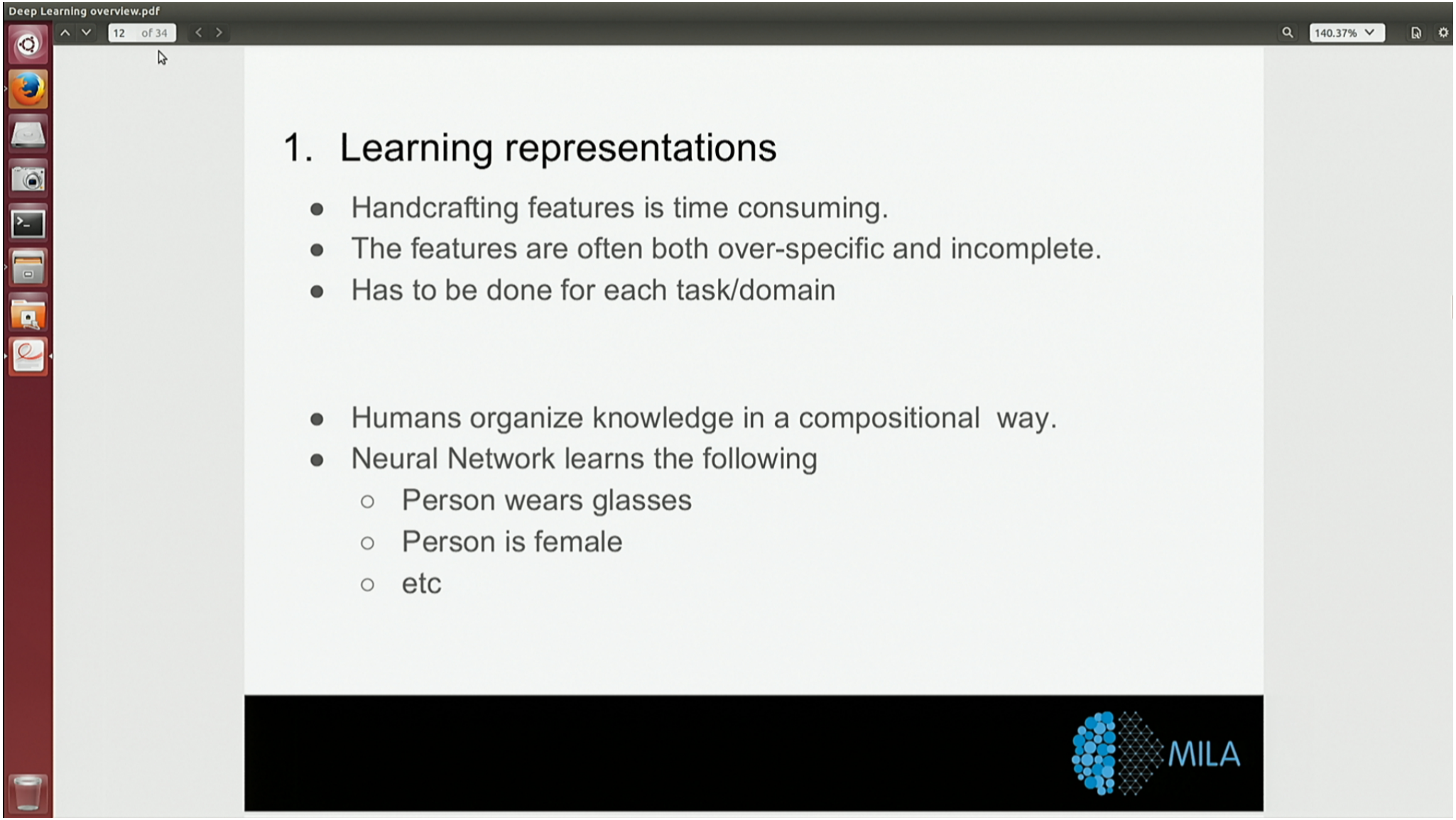
Geometrical view on machine learning

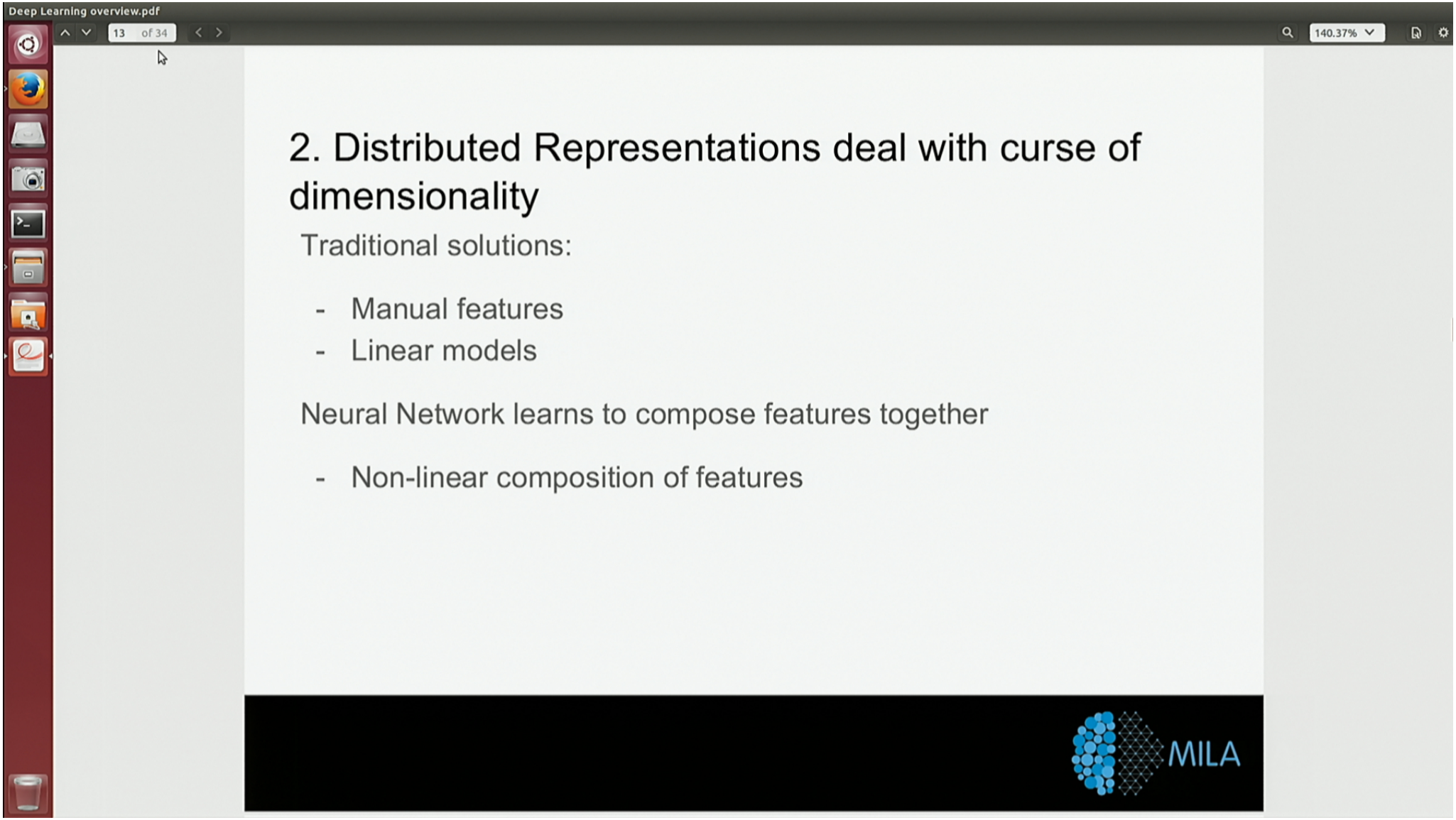
- Generalization: Guessing **where** *probability* mass concentrates.
- Challenge: The curse of dimensionality (exponentially many configurations of the variables to consider)

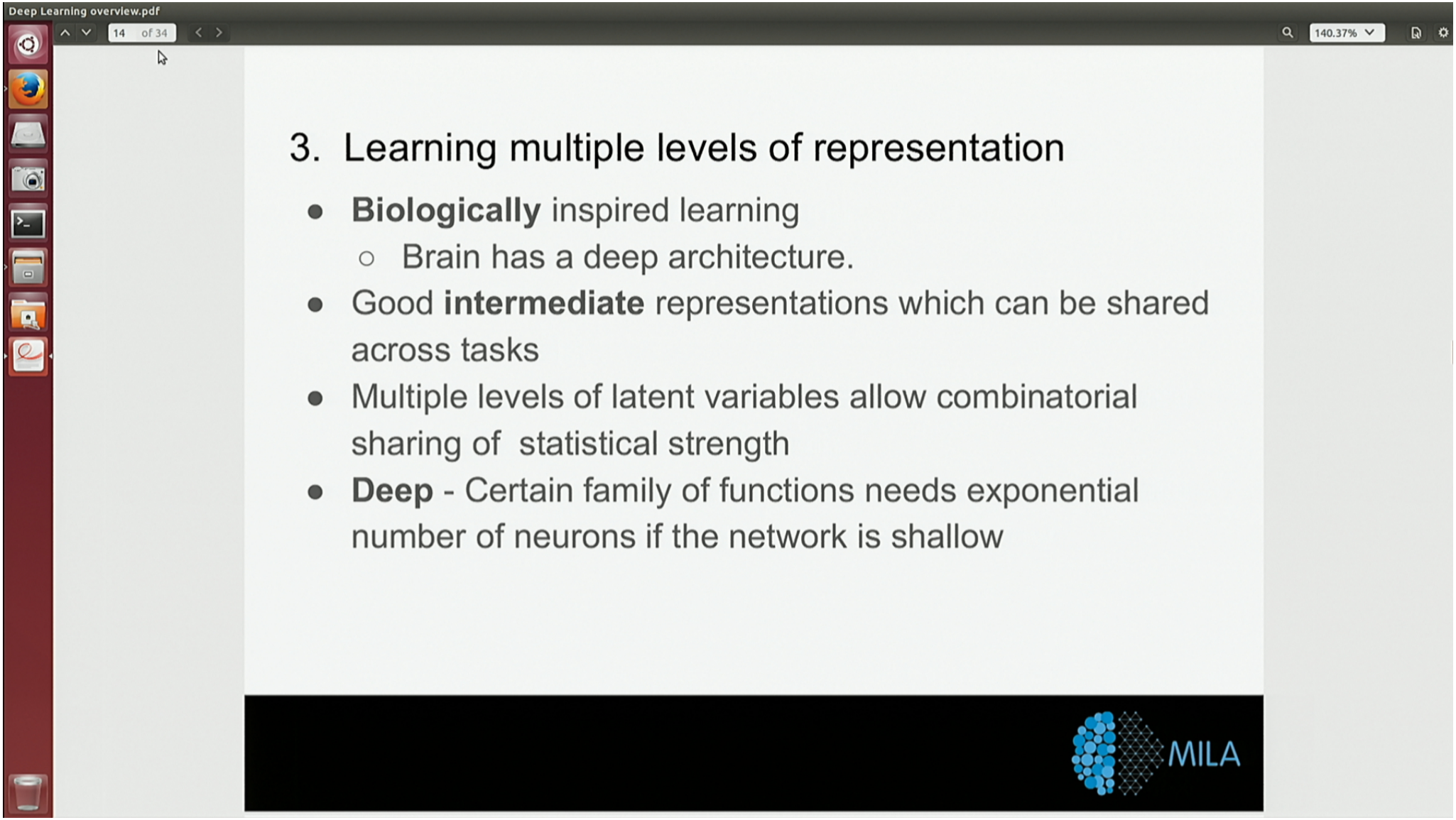


MILA


Deep Learning overview.pdf
11 of 34
140.37%
Why Deep Learning?
MILA


Deep Learning overview.pdf
12 of 34
140.37%
1. Learning representations
Handcrafting features is time consuming.
The features are often both over-specific and incomplete.
Has to be done for each task/domain
Humans organize knowledge in a compositional way.
Neural Network learns the following
Person wears glasses
Person is female
etc
MILA


Deep Learning overview.pdf
13 of 34
140.37%
2. Distributed Representations deal with curse of dimensionality
Traditional solutions:
- Manual features
- Linear models
Neural Network learns to compose features together
- Non-linear composition of features
MILA

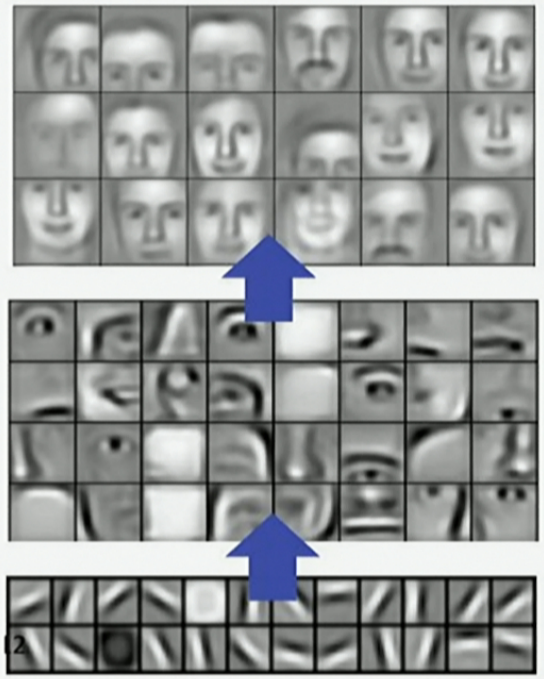

Deep Learning overview.pdf
14 of 34
140.37%
3. Learning multiple levels of representation
Biologically inspired learning
Brain has a deep architecture.
Good intermediate representations which can be shared across tasks
Multiple levels of latent variables allow combinatorial sharing of statistical strength
Deep - Certain family of functions needs exponential number of neurons if the network is shallow
MILA

Deep Learning overview.pdf

15 of 34

140.37%

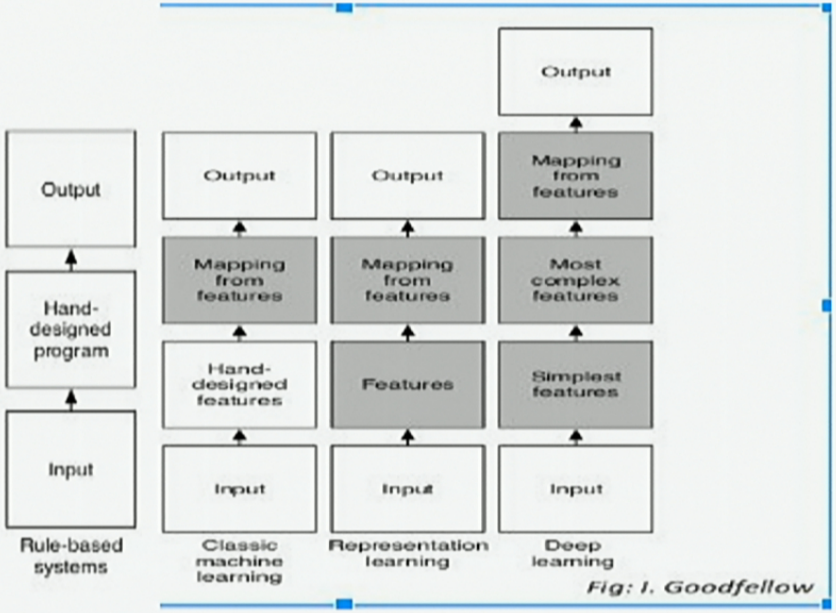
Learning multiple level of representation

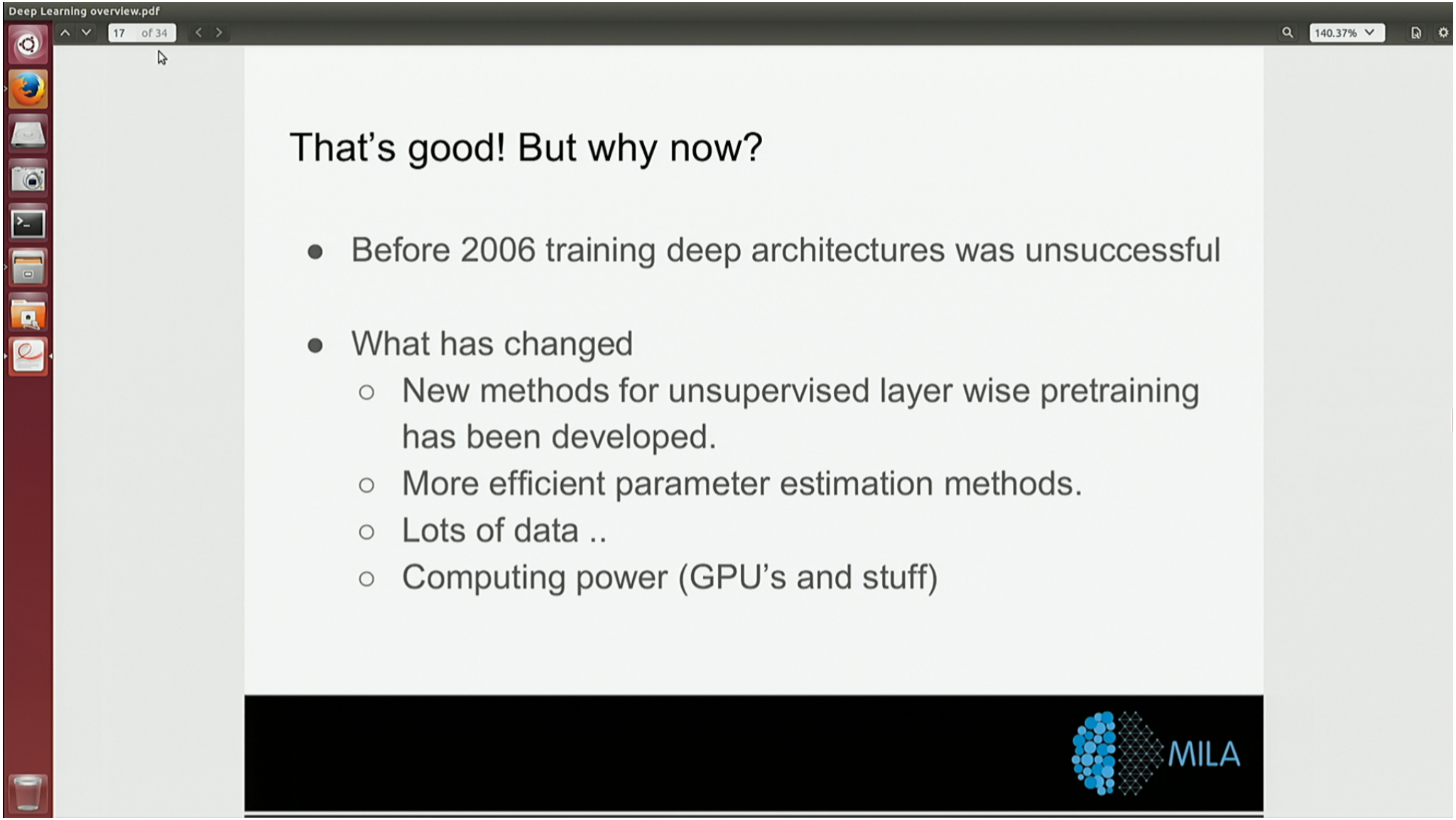


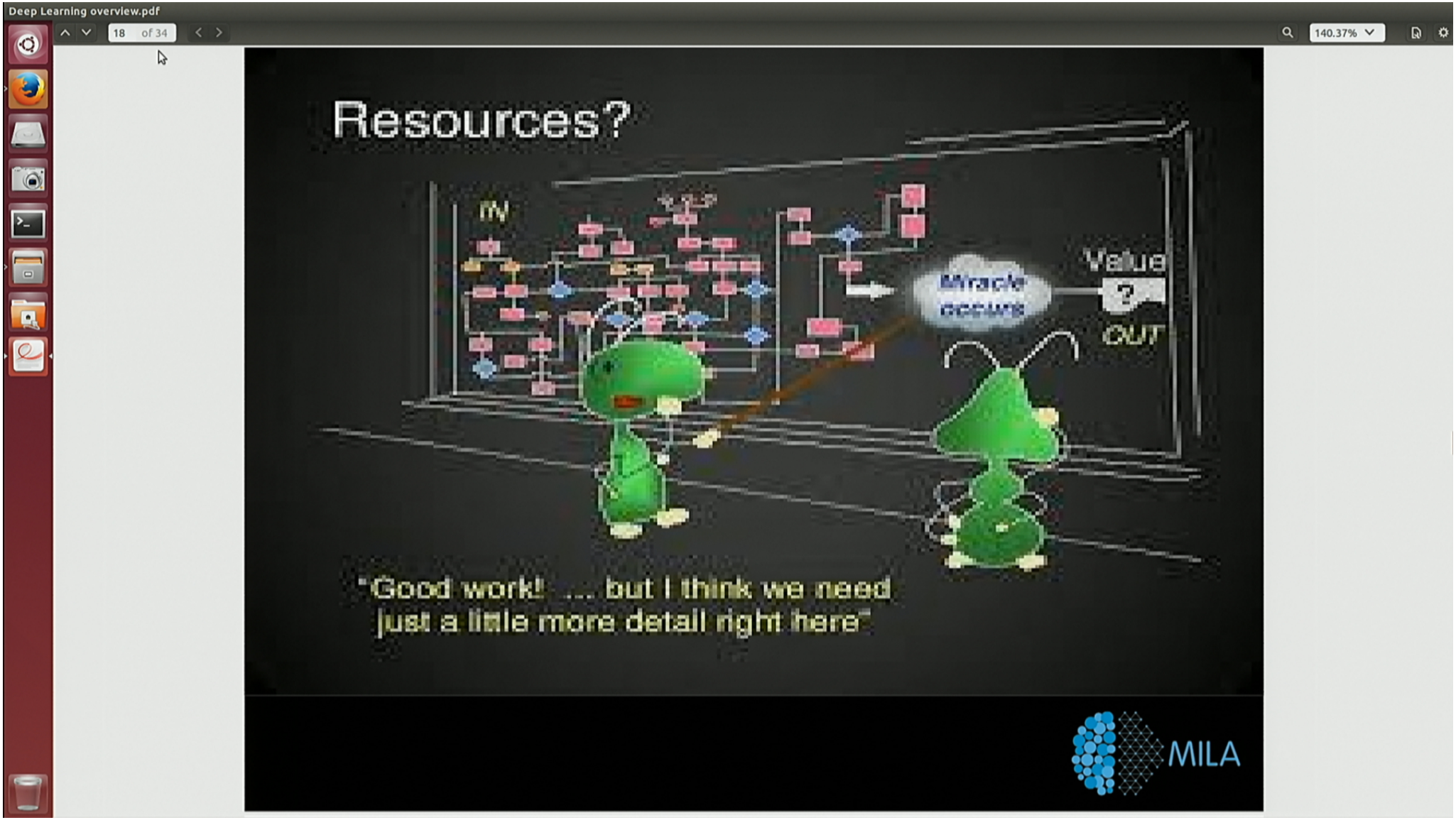
-Lee et al [2009]

MILA

Summary of different learning methods

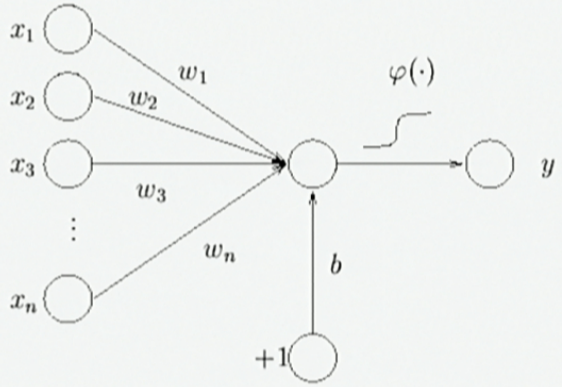



Deep Learning overview.pdf
17 of 34
140.37%
That's good! But why now?
Before 2006 training deep architectures was unsuccessful
What has changed
New methods for unsupervised layer wise pretraining has been developed.
More efficient parameter estimation methods.
Lots of data ..
Computing power (GPU's and stuff)
MILA


Deep Learning overview.pdf
18 of 34
140.37%
Resources?
IN
Miracle occurs
Value
?
OUT
"Good work! ... but I think we need just a little more detail right here"
MILA

Demystifying neural networks

- Neural Networks come with their own terminological baggage (just like SVM's)
- If you understand how logistic regression works, Then ***you already understand*** the operation of a basic neural network neuron!



MILA

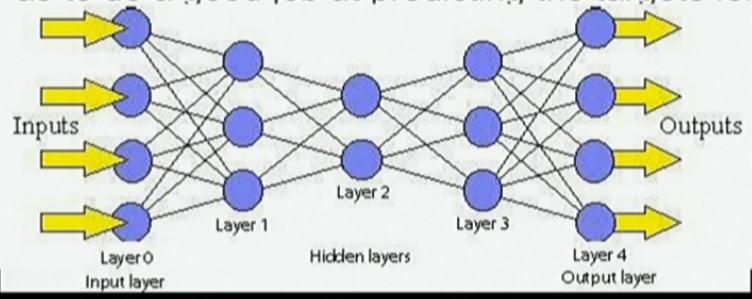
A neural network = running several logistic regressions at the same time

Feed a vector of inputs through a bunch of logistic regression functions, then we get a vector of outputs …

"But we don't have to decide ahead of time what variables these logistic regressions are trying to predict!"

Isn't it cool ?

It is the training criterion that will direct what the intermediate hidden variables should be, so as to do a good job at predicting the targets for the next layer, etc.



MILA

How to train parameters of the network

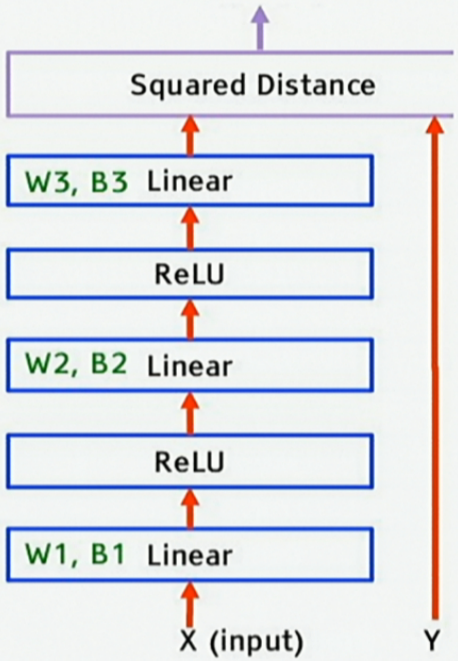
A **single** supervised layer

- Compute error derivatives (gradients)

A **multilayer** net is more complex

- Internal (“hidden”) non-linear units --> function non-convex
- We “**backpropagate**” error derivatives through the model

Feedforward Network



Objective function, Cost

Linear Module

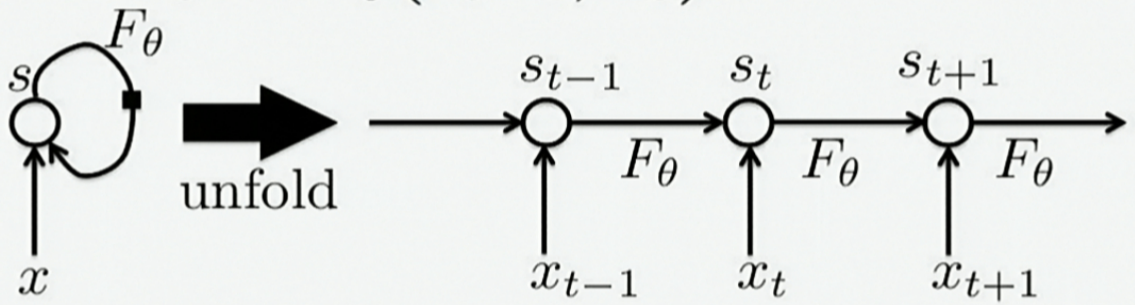
Non-Linear Module

MILA

Recurrent Neural Networks

- Shared weights across different time step

$$s_t = F_\theta(s_{t-1}, x_t)$$

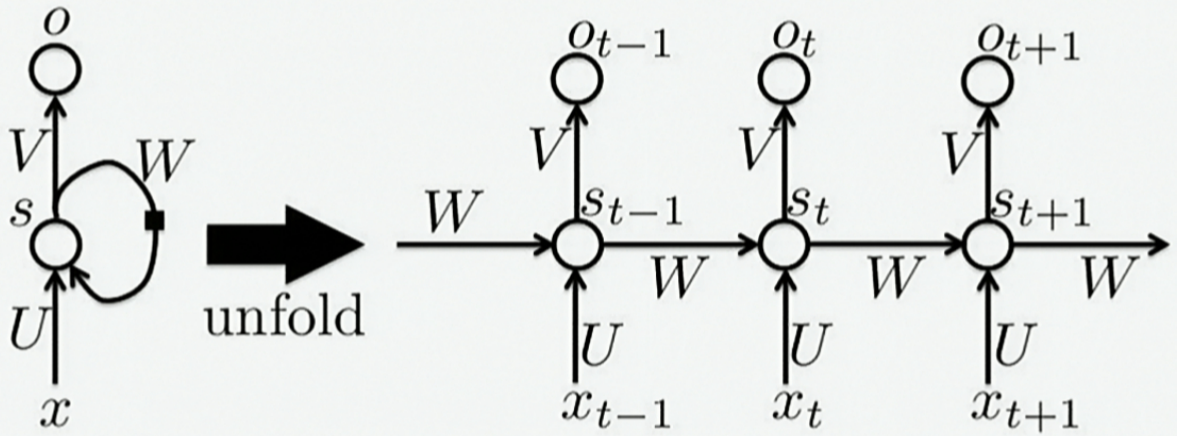


$$s_t = G_t(x_t, x_{t-1}, x_{t-2}, \ldots, x_2, x_1)$$

MILA

Recurrent Neural Networks

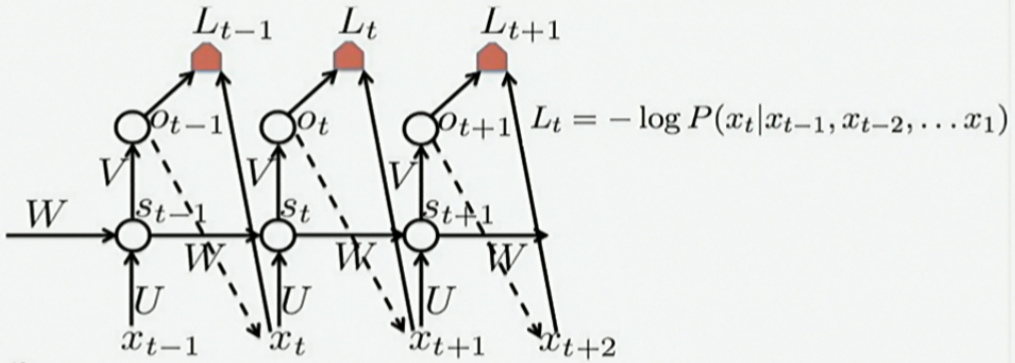
Unfold RNN graph



MILA

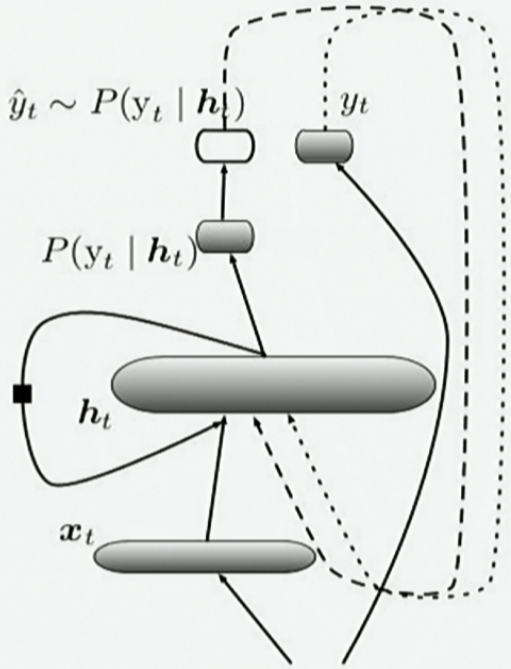
Generative RNN

$$P(\mathbf{x}) = P(x_1, \dots x_T) = \prod_{t=1}^{T} P(x_t | x_{t-1}, x_{t-2}, \dots x_1)$$



40

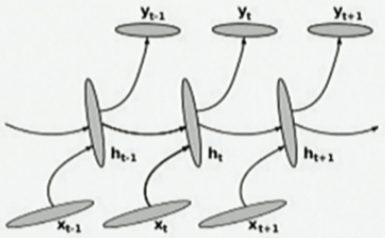
Teacher Forcing - Maximum log likelihood!



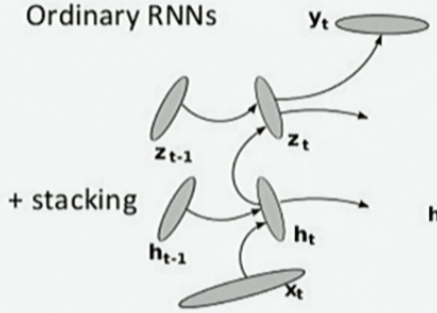
MILA

Deeper RNNs

How to construct deeper networks

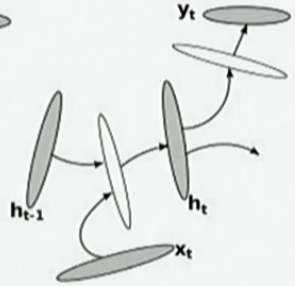


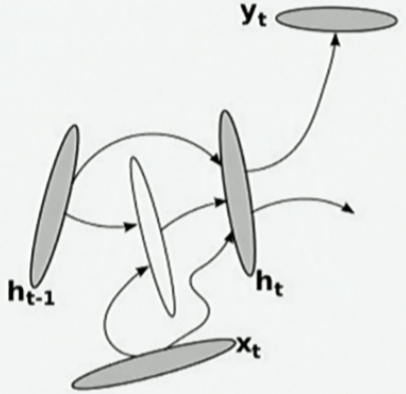
Ordinary RNNs



+ stacking

+ deep hid-to-out
+ deep hid-to-hid
+deep in-to-hid





+ skip connections for creating shorter paths

-Y.Bengio

MILA

Long term dependency

RNN gradient is a product of Jacobians

$$L = L(s_T(s_{T-1}(\dots s_{t+1}(s_t, \dots))))$$

$$\frac{\partial L}{\partial s_t} = \frac{\partial L}{\partial s_T}\frac{\partial s_T}{\partial s_{T-1}} \cdots \frac{\partial s_{t+1}}{\partial s_t}$$

Storing bits robustly requires sing. values<1

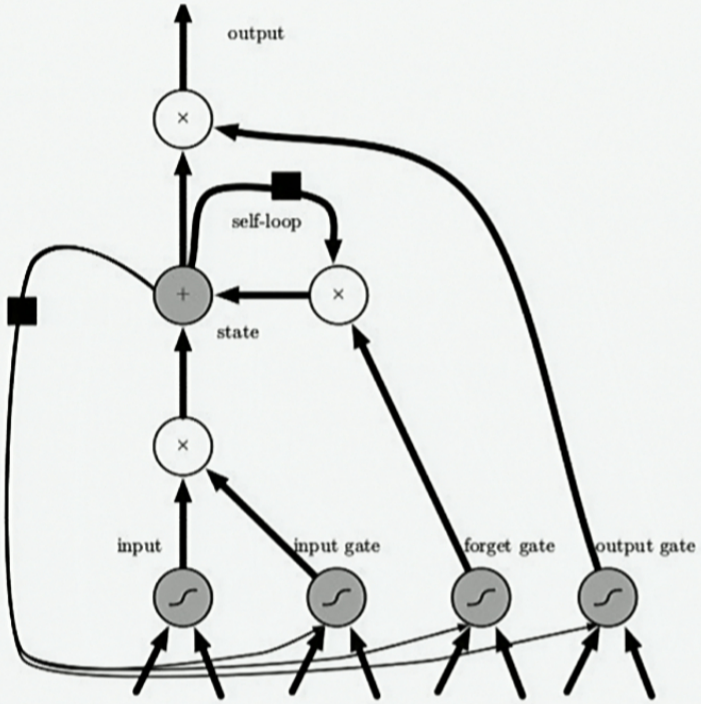
Eigenvalue of Jacobian > 1 → exploding gradient - gradient clipping

Eigenvalue of Jacobian < 1 → vanishing gradient

MILA

GRU and LSTM

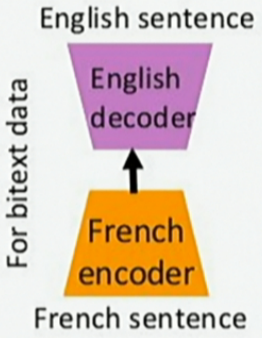
- Path where gradients can flow for longer
- Eigenvalue of Jacobian slightly less than 1
- Long Short Term Memory
- Gated Recurrent Networks (lighter version)

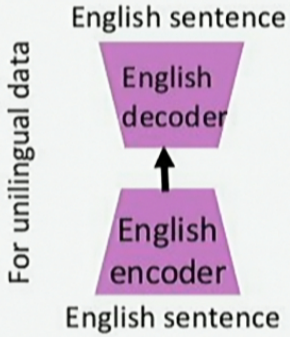


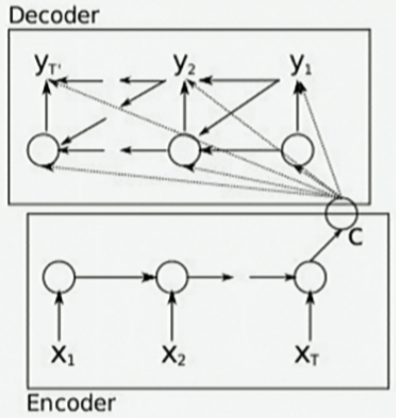
MILA

Encoder-Decoder Network

- Intermediate representation of meaning = 'universal representation'
- Encoder: word sequence → sentence representation
- Decoder: sentence representation → word sequence





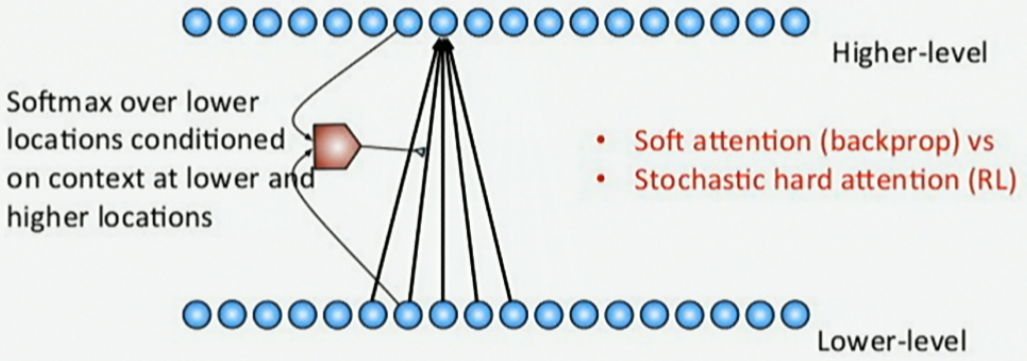


Cho et al EMNLP 2014; Sutskever et al NIPS 2014

MILA

Attention Model for Neural Networks

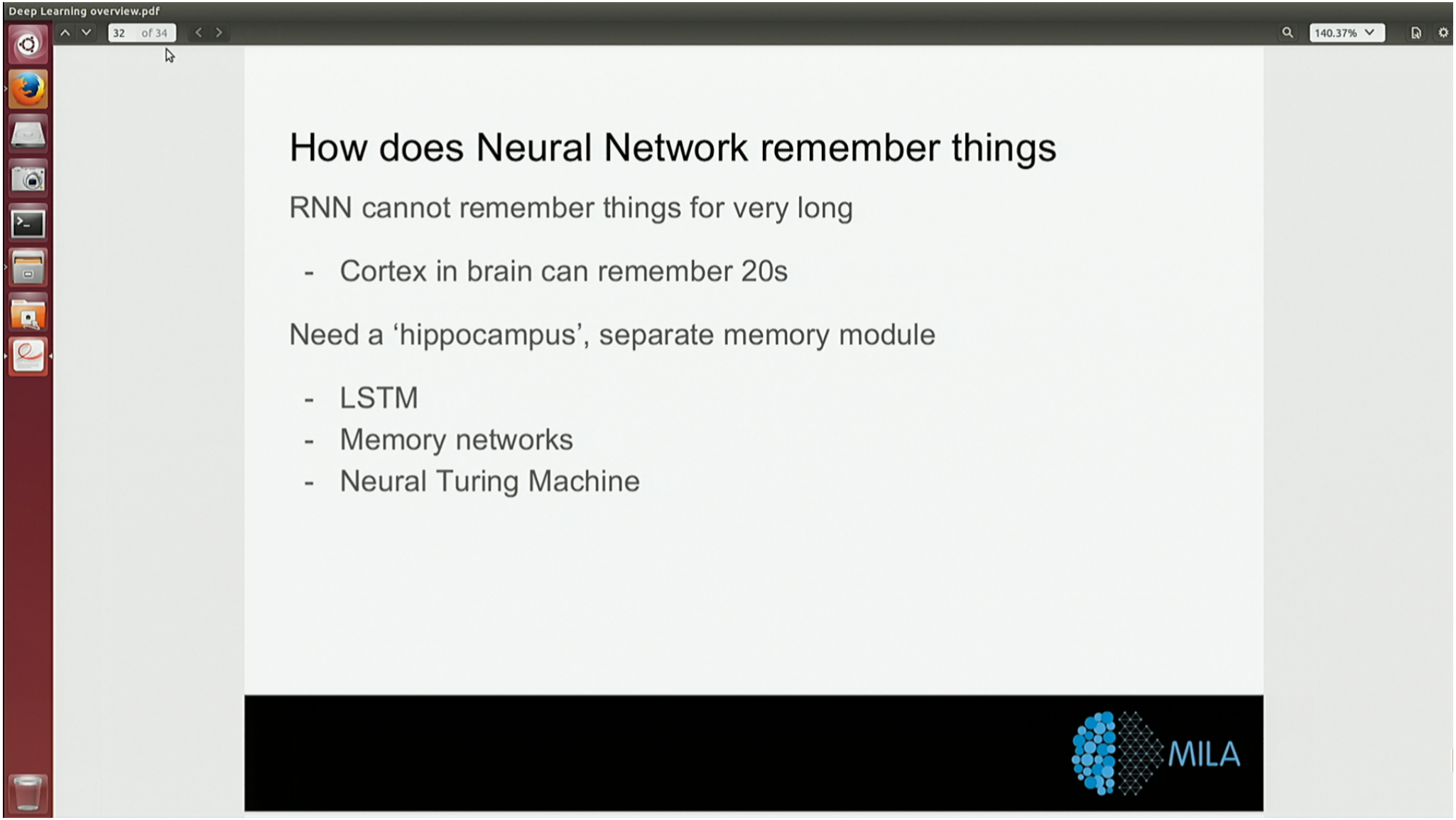
- Sequence or image as input (or immediate)
- Upper level representation chooses where to look
  - Assigning a weight (probability) to each input position



- Soft attention (backprop) vs
- Stochastic hard attention (RL)

Bahdanau, Cho & Bengio, arXiv sept. 2014

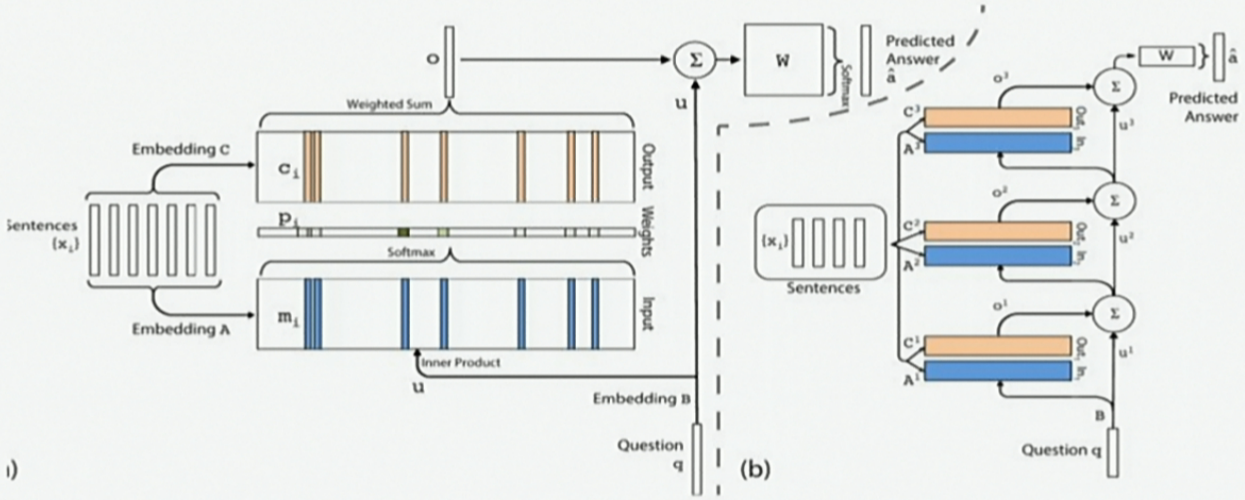
MILA


Deep Learning overview.pdf
32 of 34
140.37%
How does Neural Network remember things
RNN cannot remember things for very long
- Cortex in brain can remember 20s
Need a ‘hippocampus’, separate memory module
- LSTM
- Memory networks
- Neural Turing Machine
MILA

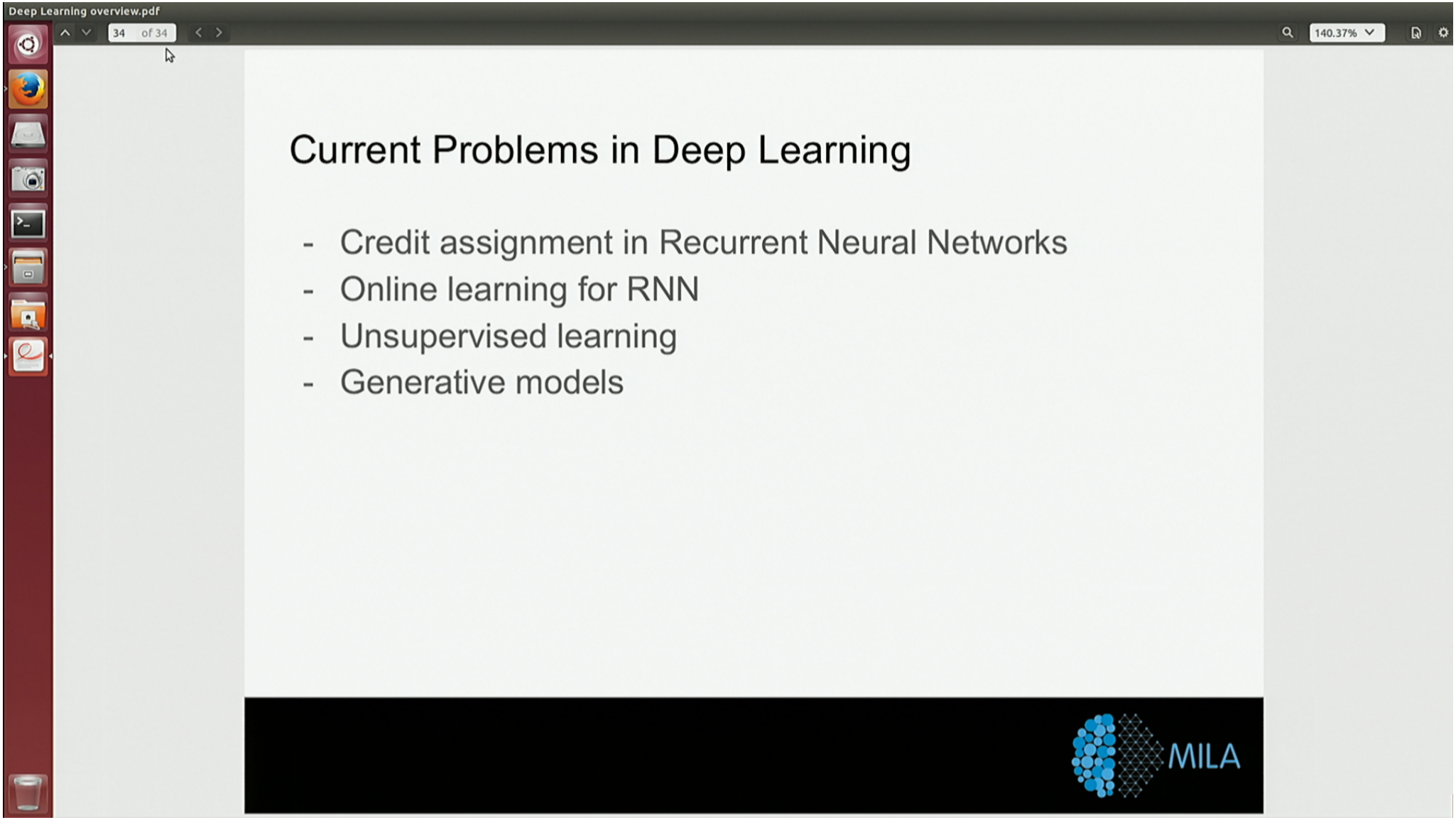
End-To-End Memory Networks

Weakly-supervised MemNN

- no need to tell which memory location to use.



Sukhbataar, Szlam, Weston, Fergus NIPS 2015, ArXiv:1503.08895]


Deep Learning overview.pdf
34 of 34
140.37%
Current Problems in Deep Learning
- Credit assignment in Recurrent Neural Networks
- Online learning for RNN
- Unsupervised learning
- Generative models
MILA