Introduction to Deep Learning for Text Analysis and Understanding

COSC 7336: Advanced Natural Language Processing
Fall 2017

UNIVERSITY of HOUSTON

UNIVERSIDAD NACIONAL DE COLOMBIA

Instructors

	Fabio Gonzalez Full Professor UNIVERSIDAD NACIONAL DE COLOMBIA National University of Colombia Visiting Professor at UH Email: fagonzalezo@unal.edu.co Office: PGH 598
	Thamar Solorio Associate Professor UNIVERSITY of HOUSTON University of Houston Email: thamar.solorio@gmail.com Office: PGH 584

Today's Lecture

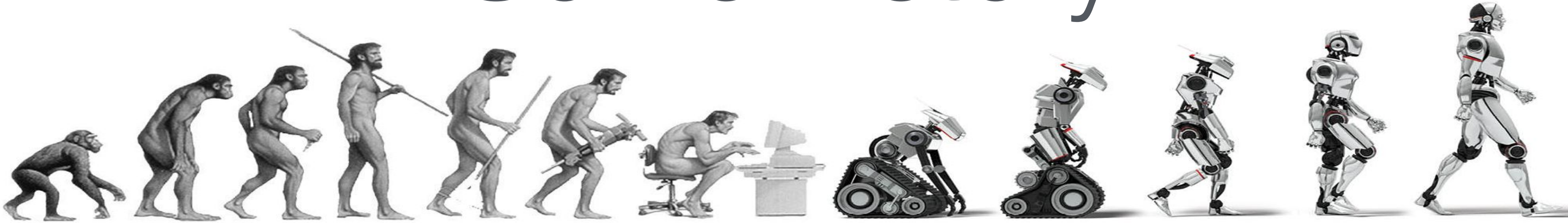
★ Intro to DL
★ Why DL is a promising direction to solve NLP problems
★ Overview of the field of NLP
★ Course Administrivia

Intro to DL

UNIVERSITY of HOUSTON

UNIVERSIDAD NACIONAL DE COLOMBIA

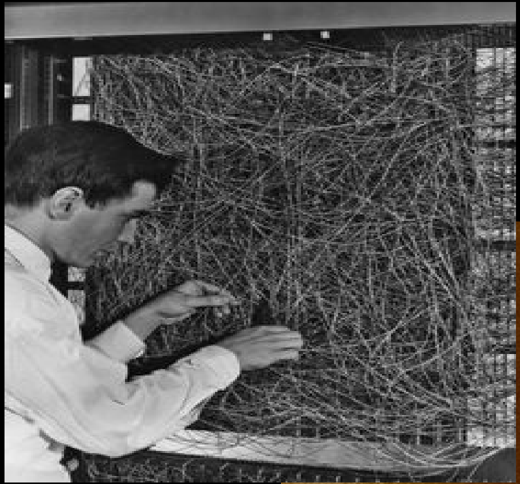
Some history

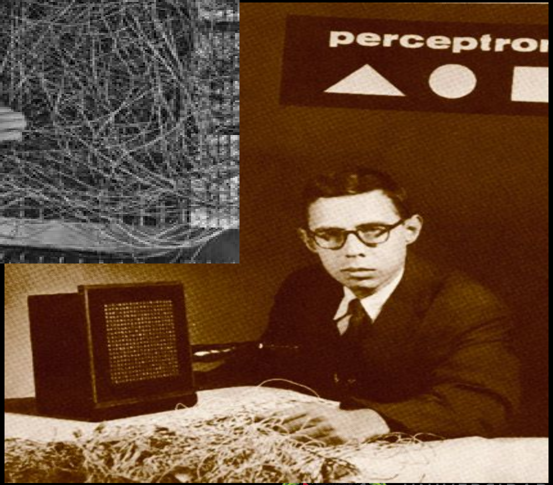


https://www.youtube.com/watch?v=cNxadbrN_aI

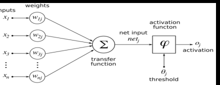
Rosenblatt’s Perceptron (1957)

- Input: 20x20 photocells array
- Weights implemented with potentiometers
- Weight updating performed by electric motors



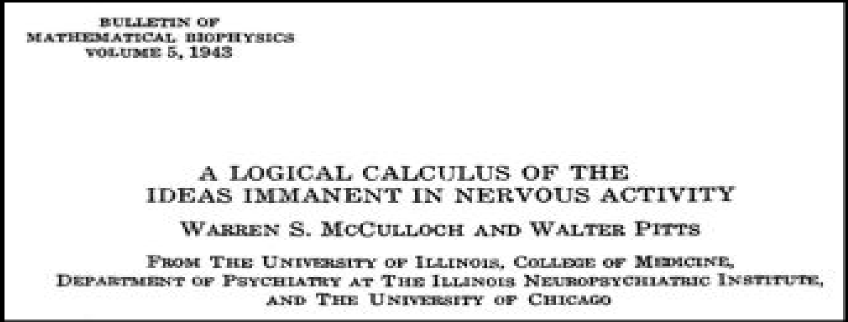


Neural networks timeline



1943 1957 1969 1986 1995 2007 2012 2017

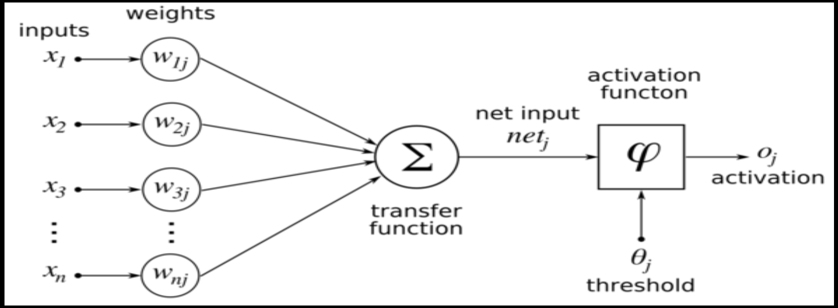
McCulloch & Pitts Artificial Neuron

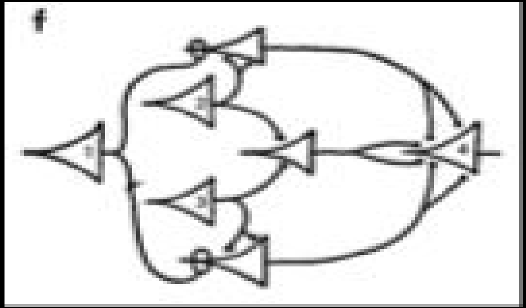

BULLETIN OF
MATHEMATICAL BIOPHYSICS
VOLUME 5, 1943

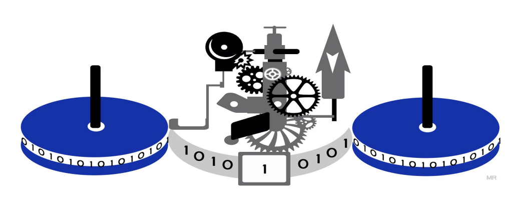
A LOGICAL CALCULUS OF THE
IDEAS IMMANENT IN NERVOUS ACTIVITY

WARREN S. MCCULLOCH AND WALTER PITTS

FROM THE UNIVERSITY OF ILLINOIS, COLLEGE OF MEDICINE,
DEPARTMENT OF PSYCHIATRY AT THE ILLINOIS NEUROPSYCHIATRIC INSTITUTE,
AND THE UNIVERSITY OF CHICAGO



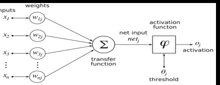


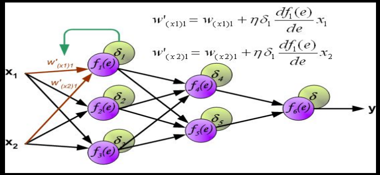


UNIVERSITY of
HOUSTON

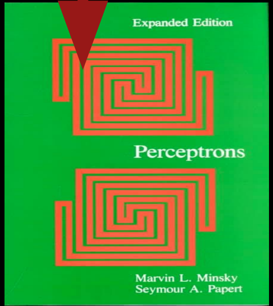
UNIVERSIDAD
NACIONAL
DE COLOMBIA

Neural networks timeline

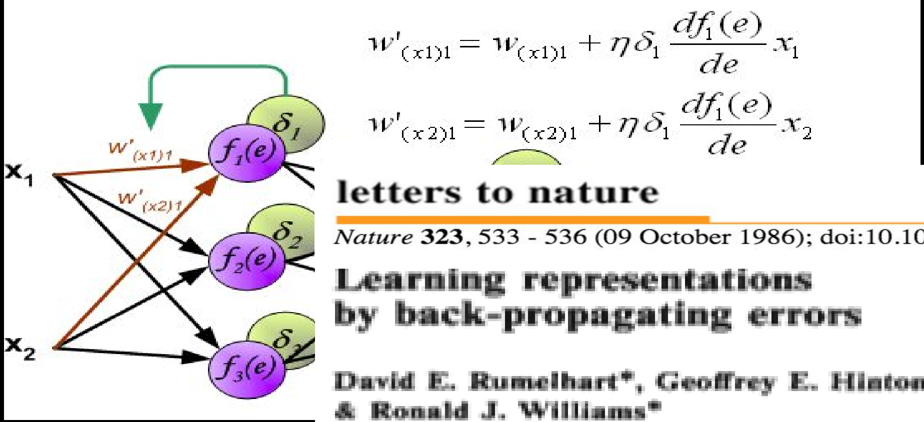




1943 1957 1969 1986 1995 2007 2012 2017



Backpropagation



$$w'_{(x1)1} = w_{(x1)1} + \eta \delta_1 \frac{df_1(e)}{de} x_1$$

$$w'_{(x2)1} = w_{(x2)1} + \eta \delta_1 \frac{df_1(e)}{de} x_2$$

letters to nature

Nature **323**, 533 - 536 (09 October 1986); doi:10.1038/323533a0

nature

Learning representations by back-propagating errors

David E. Rumelhart*, Geoffrey E. Hinton† & Ronald J. Williams*

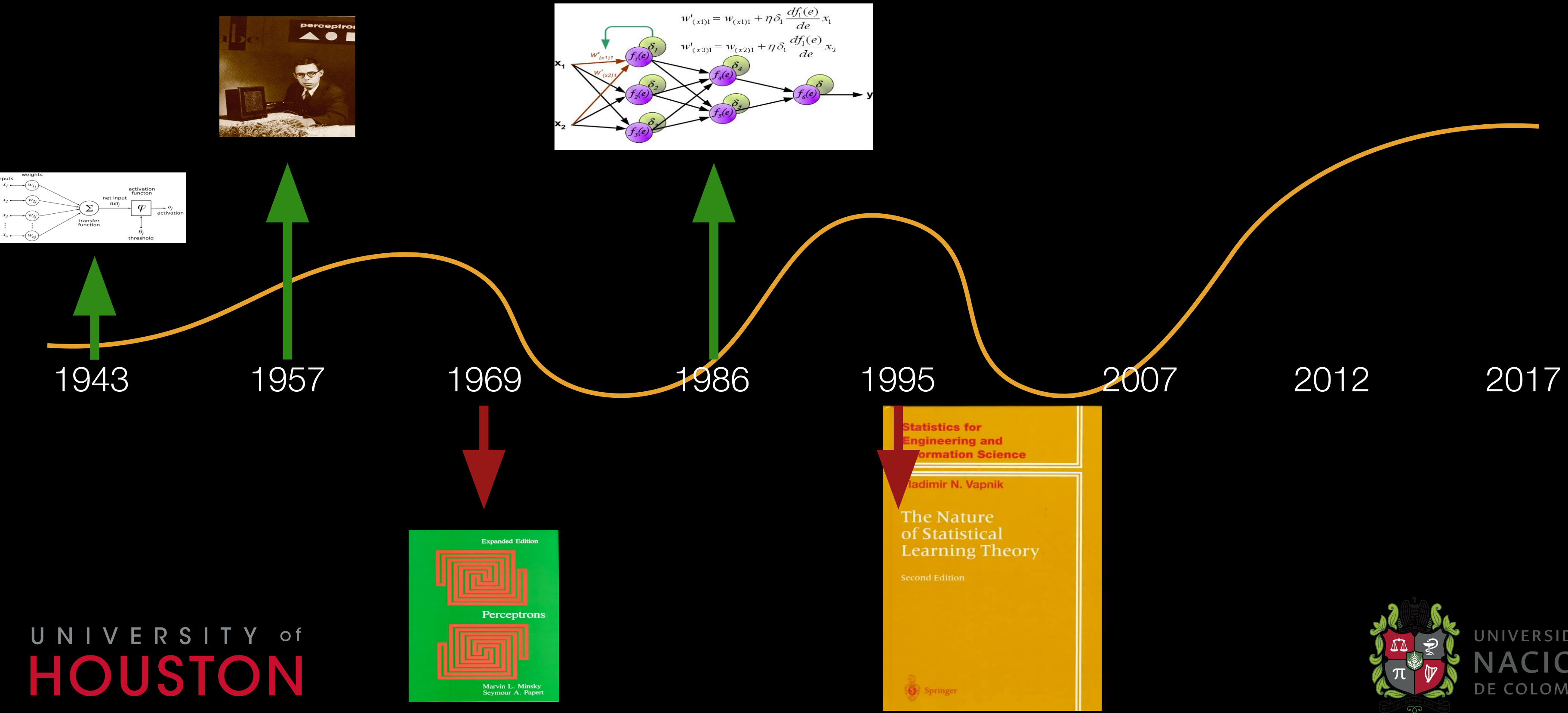
* Institute for Cognitive Science, C-015, University of Californi
San Diego, La Jolla, California 92093, USA
† Department of Computer Science, Carnegie-Mellon Universit
Pittsburgh, Philadelphia 15213, USA

Source: http://home.agh.edu.pl/~vlsi/AI/backp_t_en/backprop.html

UNIVERSITY of HOUSTON

UNIVERSIDAD NACIONAL DE COLOMBIA

Neural networks timeline



UNIVERSITY of HOUSTON

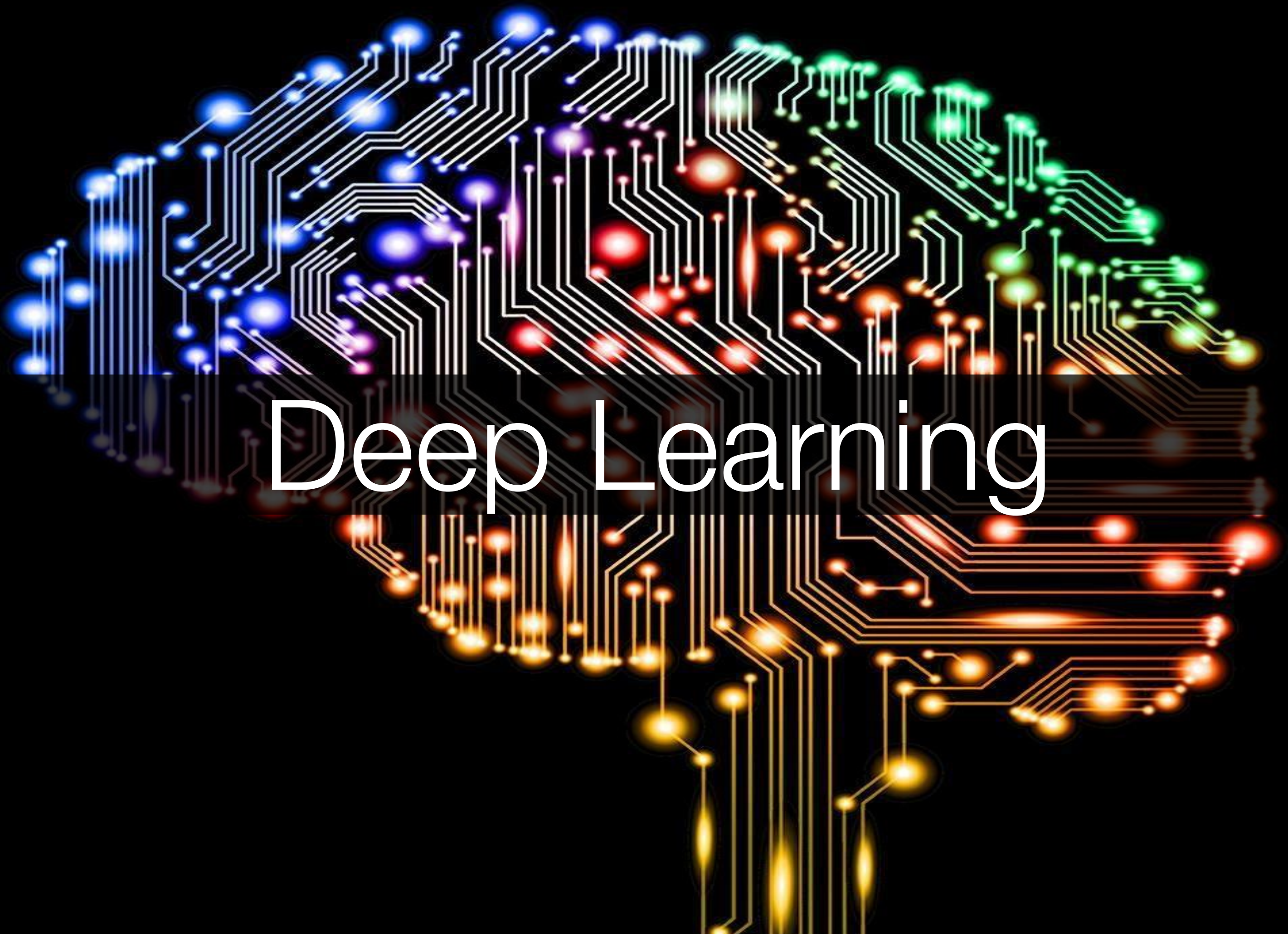
UNIVERSIDAD NACIONAL DE COLOMBIA

My own history with NN (circa 1993)

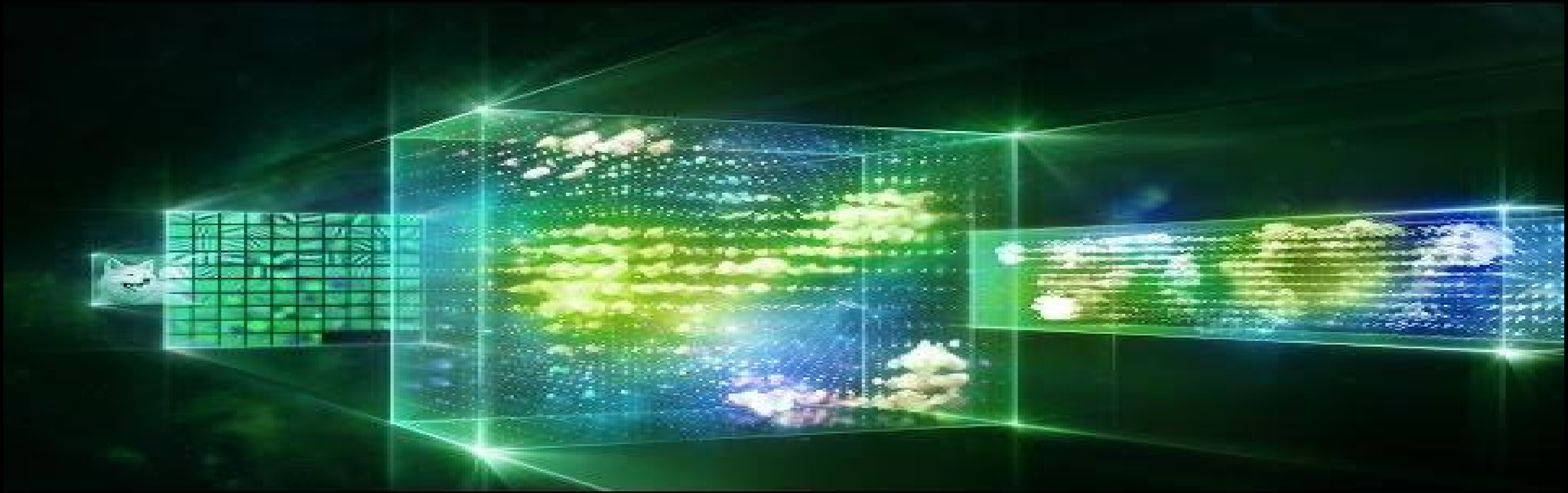
Quick and Dirty Introduction to Neural Networks

Interactive Demo

Deep Learning

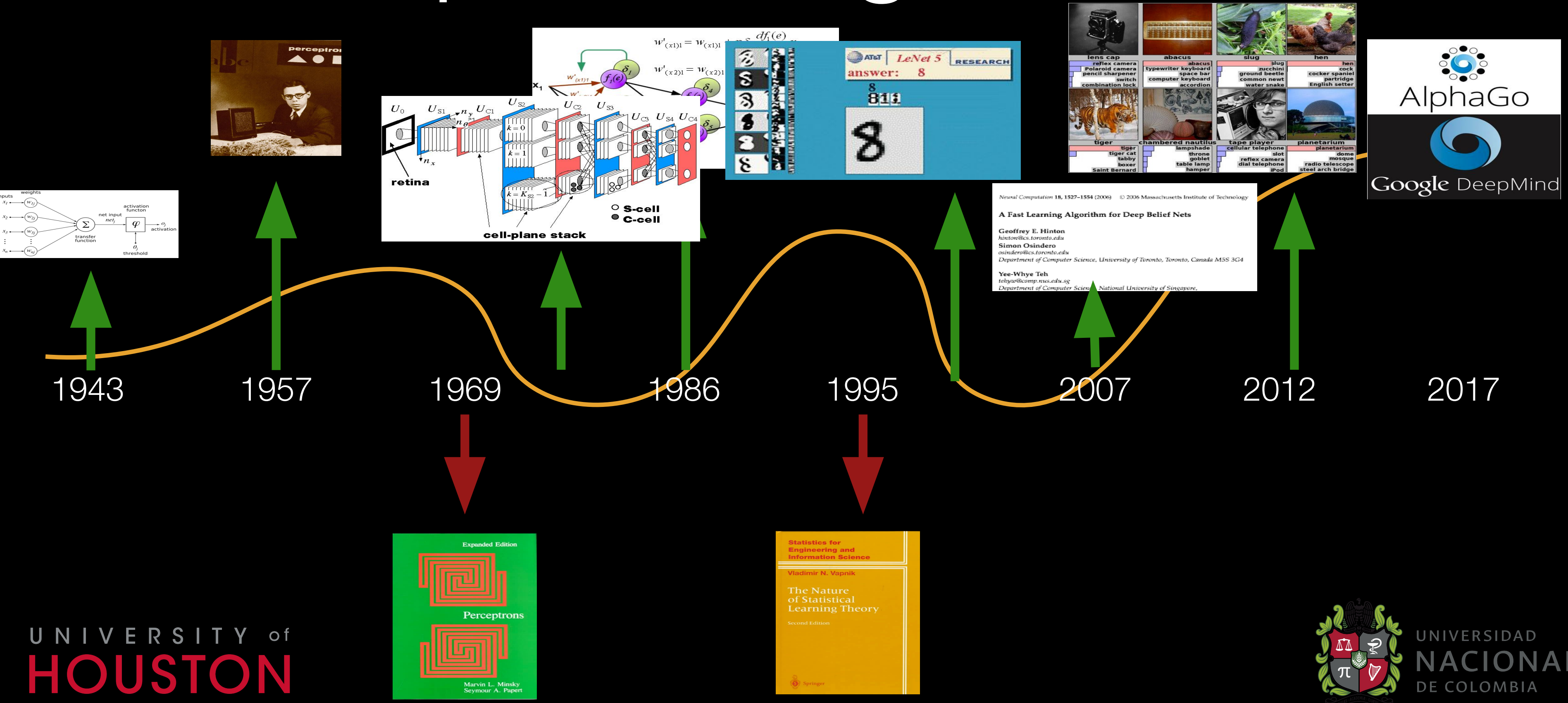


Deep learning boom



Deep learning boom

Deep learning time line



UNIVERSITY of HOUSTON

UNIVERSIDAD NACIONAL DE COLOMBIA

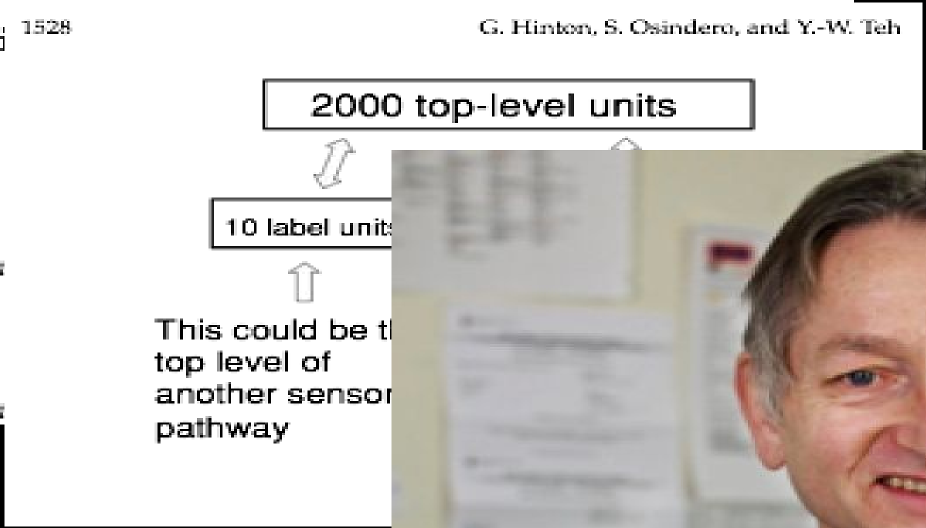
Deep Learning is Born

Neural Computation 18, 1527–1554 (2006)

A Fast Learning Algori

Geoffrey E. Hinton
hinton@cs.toronto.edu
Simon Osindero
osindero@cs.toronto.edu
Department of Computer Scien

Yee-Whye Teh
tehyw@comp.nus.edu.sg
Department of Computer Scien



Deep learning model won ILSVRC 2012 challenge

- ILSV… Larg… Reco…

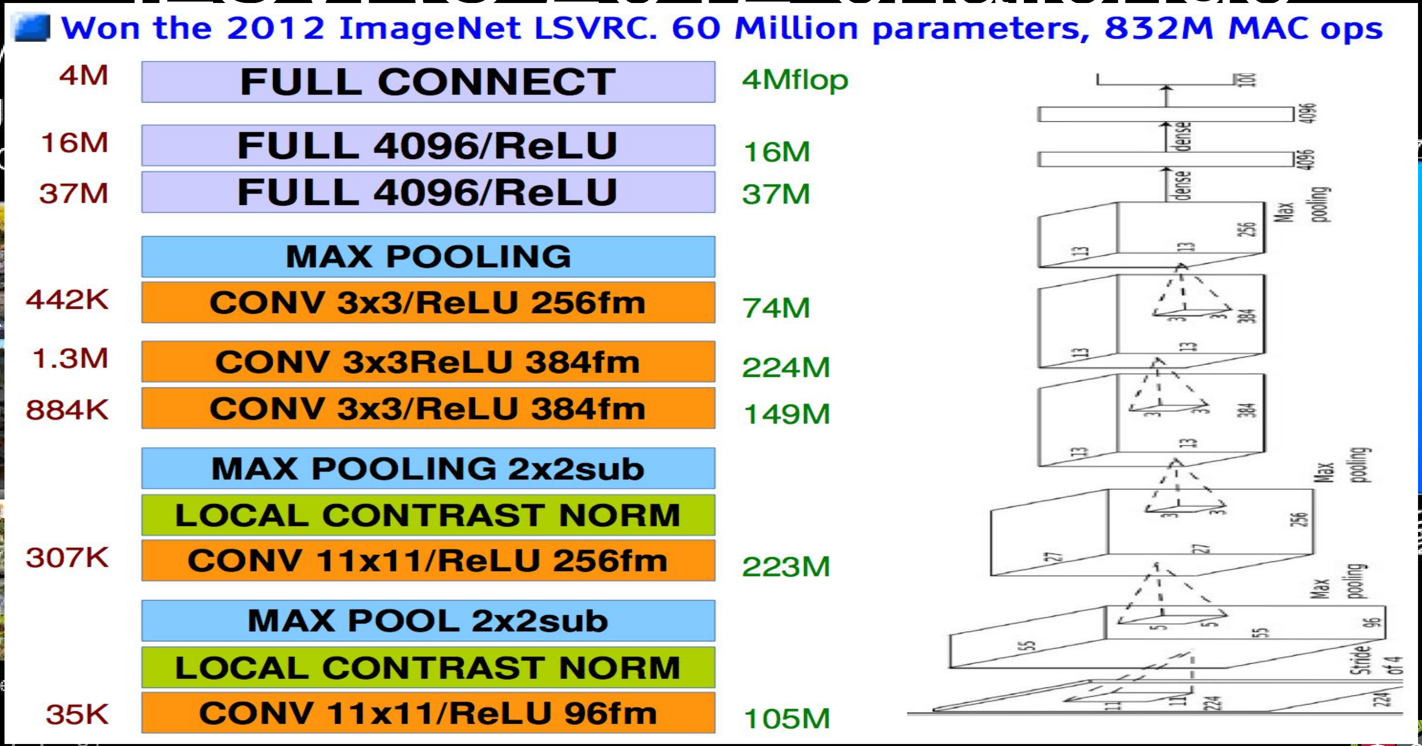
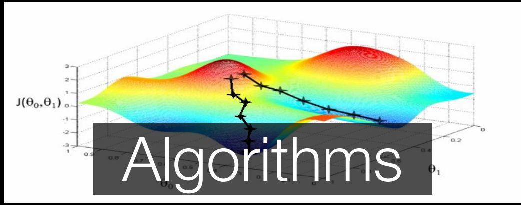


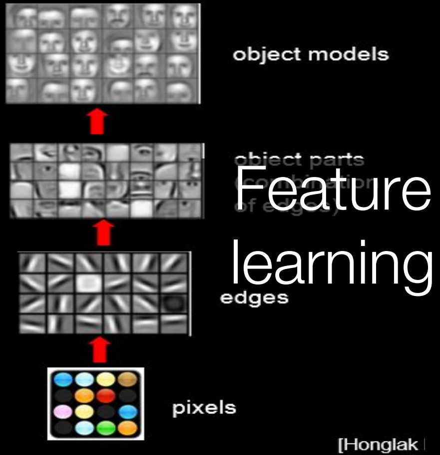
Image source…

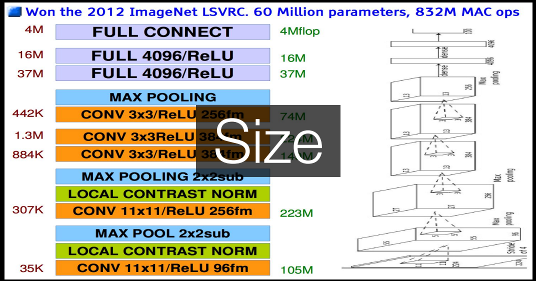
76%

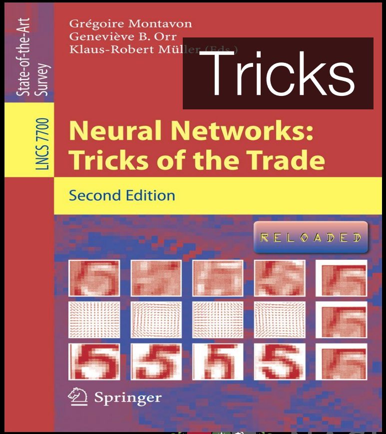
Univ of Amsterdam

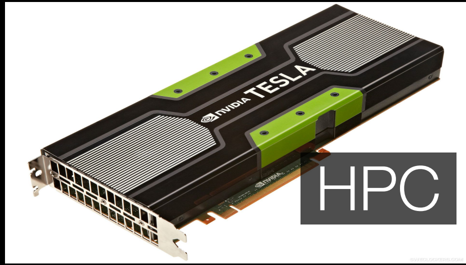
Deep learning recipe



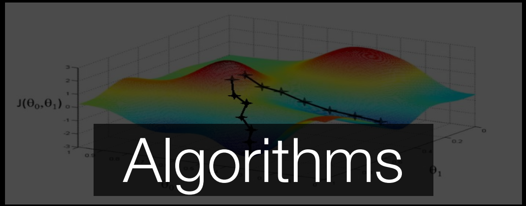


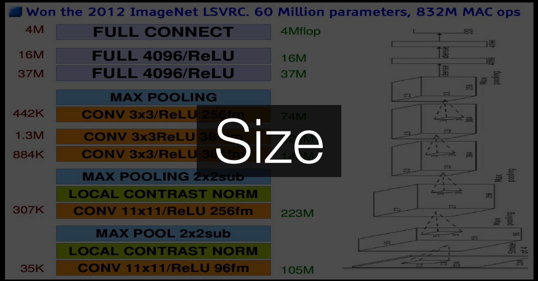


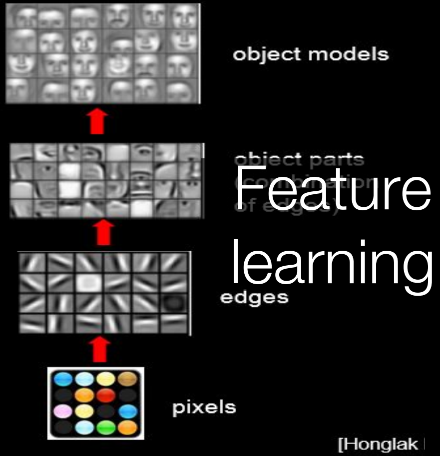


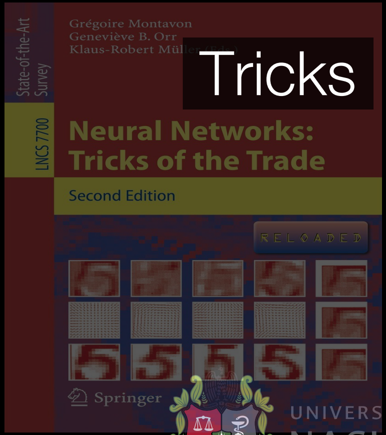


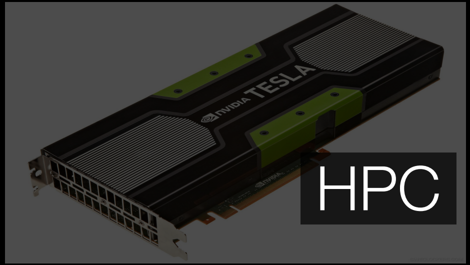
Deep learning recipe







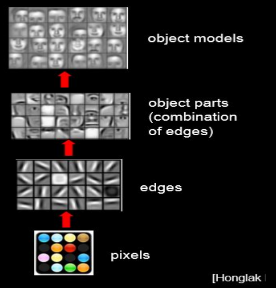


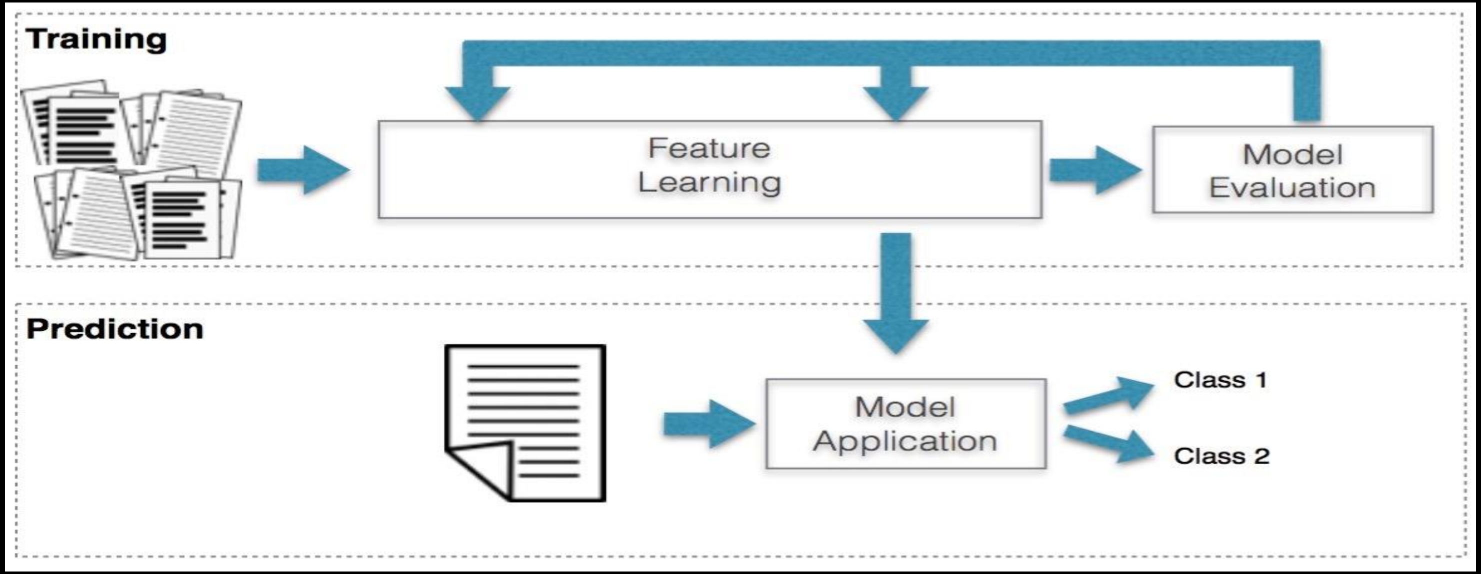


UNIV HOUSTON

UNIVERSIDAD NACIONAL DE COLOMBIA

Feature learning

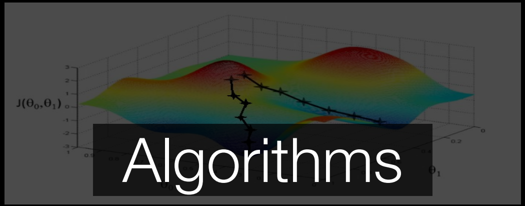


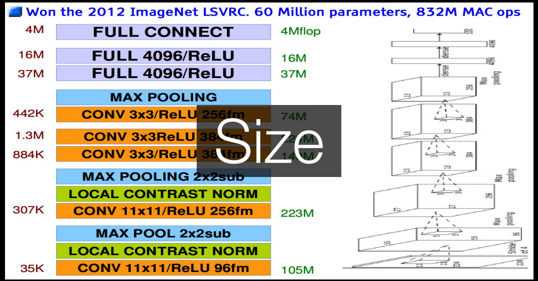


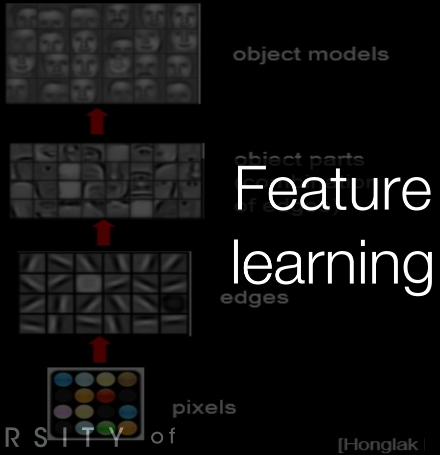
UNIVERSITY of HOUSTON

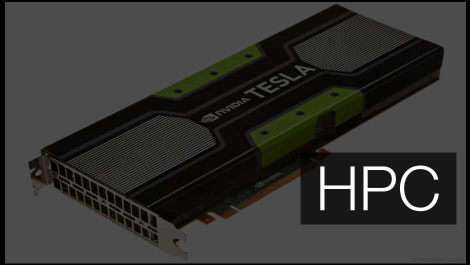
UNIVERSIDAD NACIONAL DE COLOMBIA

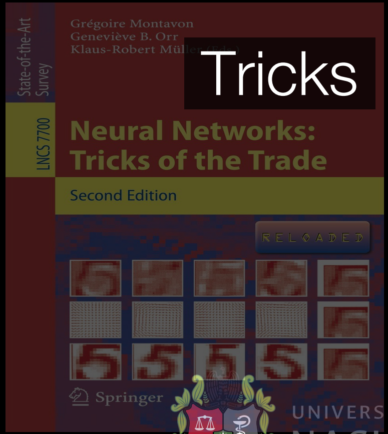
Deep learning recipe







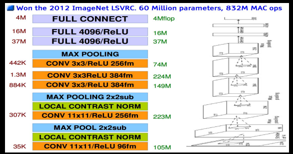




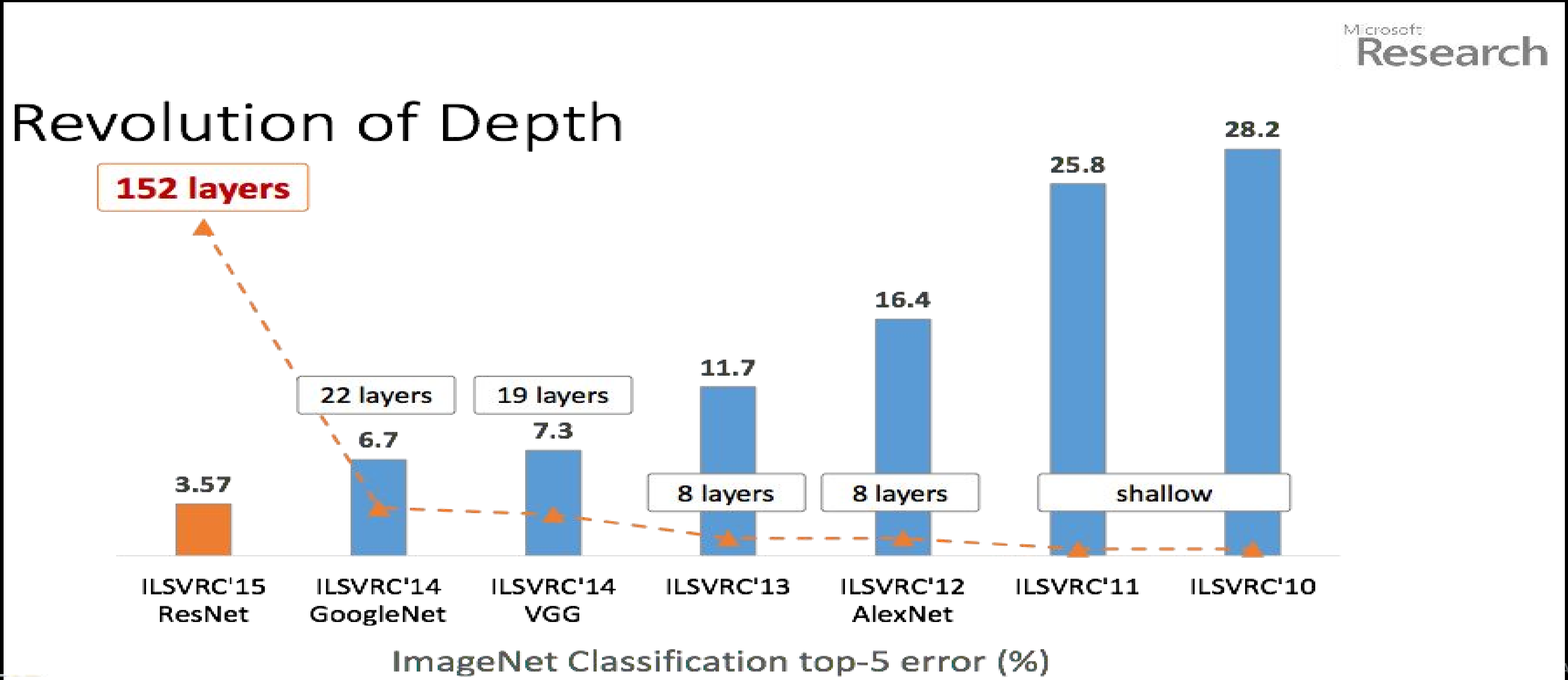
UNIVERSITY of HOUSTON

UNIVERSIDAD NACIONAL DE COLOMBIA

Deep —> Bigger

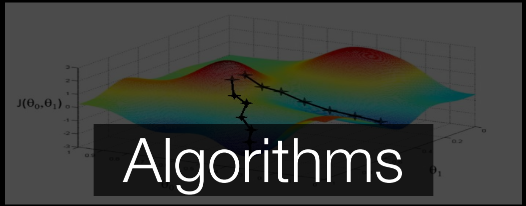


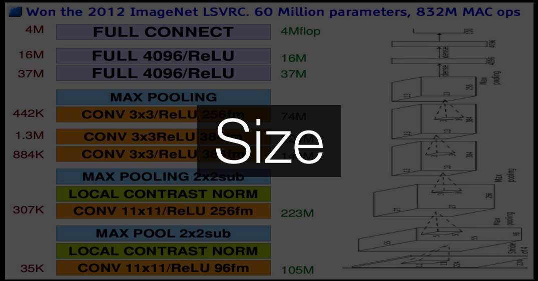
Won the 2012 ImageNet LSVRC. 60 Million parameters, 832M MAC ops
4M
FULL CONNECT
4Mflop
16M
FULL 4096/ReLU
16M
37M
FULL 4096/ReLU
37M
MAX POOLING
442K
CONV 3x3/ReLU 256fm
74M
1.3M
CONV 3x3ReLU 384fm
224M
884K
CONV 3x3/ReLU 384fm
149M
MAX POOLING 2x2sub
LOCAL CONTRAST NORM
307K
CONV 11x11/ReLU 256fm
223M
MAX POOL 2x2sub
LOCAL CONTRAST NORM
35K
CONV 11x11/ReLU 96fm
105M

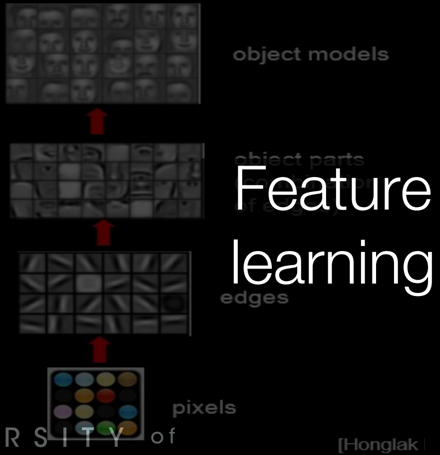


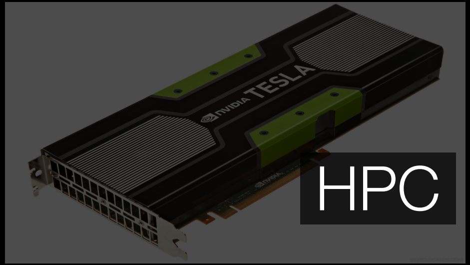
Microsoft Research
Revolution of Depth
152 layers
22 layers
19 layers
8 layers
8 layers
shallow
3.57
6.7
7.3
11.7
16.4
25.8
28.2
ILSVRC'15 ResNet
ILSVRC'14 GoogleNet
ILSVRC'14 VGG
ILSVRC'13
ILSVRC'12 AlexNet
ILSVRC'11
ILSVRC'10
ImageNet Classification top-5 error (%)

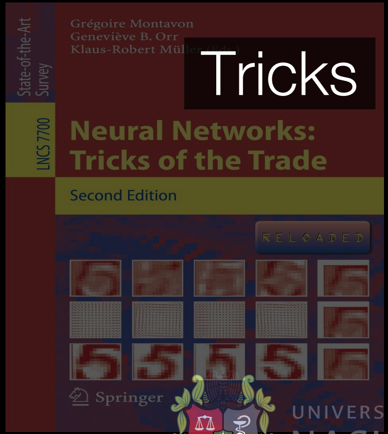
Deep learning recipe









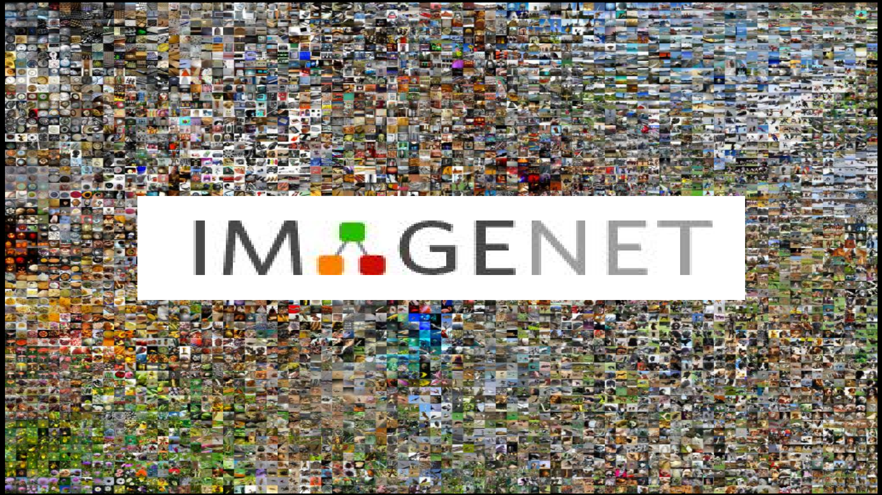


UNIVERSITY of HOUSTON

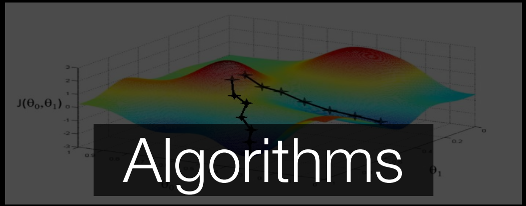
UNIVERSIDAD NACIONAL DE COLOMBIA

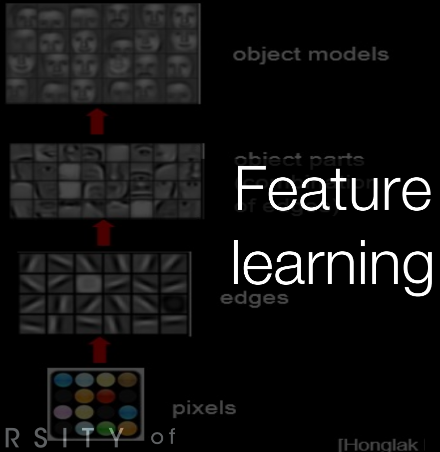
Data...

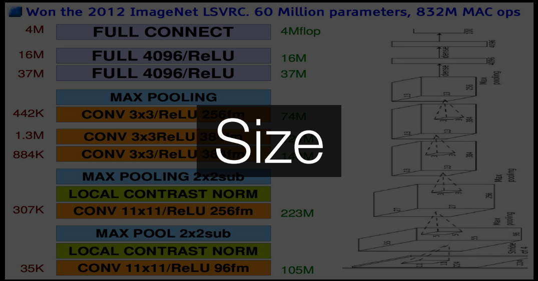
- Images annotated with WordNet concepts
- Concepts: 21,841
- Images: 14,197,122
- Bounding box annotations: 1,034,908
- Crowdsourcing

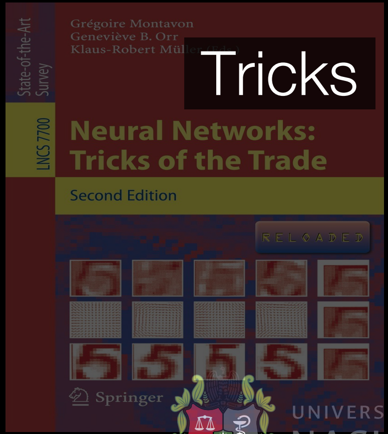


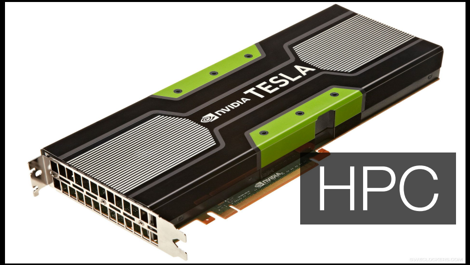
Deep learning recipe











UNIVERSITY of HOUSTON

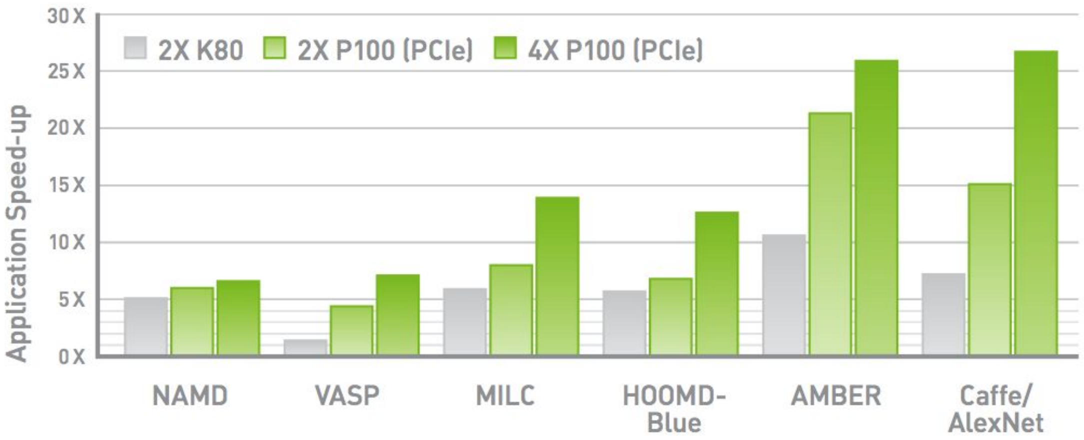
UNIVERSIDAD NACIONAL DE COLOMBIA

HPC

SPECIFICATIONS

GPU Architecture	**NVIDIA Pascal**
NVIDIA CUDA® Cores	**3584**
Double-Precision Performance	**4.7 TeraFLOPS**
Single-Precision Performance	**9.3 TeraFLOPS**
Half-Precision Performance	**18.7 TeraFLOPS**
GPU Memory	**16GB CoWoS HBM2 at 732 GB/s or 12GB CoWoS HBM2 at 549 GB/s**
System Interface	**PCIe Gen3**
Max Power Consumption	**250 W**
ECC	**Yes**
Thermal Solution	**Passive**
Form Factor	**PCIe Full Height/Length**

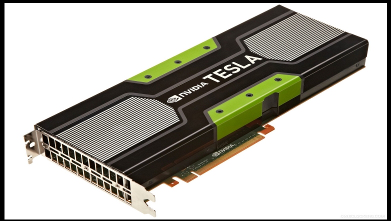
NVIDIA Tesla P100 for PCIe Performance



Dual CPU server, Intel E5-2698 v3 @ 2.3 GHz, 256 GB System Memory, Pre-Production Tesla P100

UNIVERSIDAD NACIONAL DE COLOMBIA

HPC



NVIDIA CUDA

python™

SciPy

NumPy

CONTINUUM ANALYTICS

K

theano

torch
A SCIENTIFIC COMPUTING FRAMEWORK FOR LUAJIT

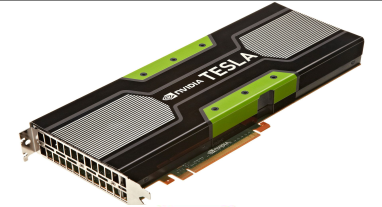
TensorFlow

Decaf / Caffe
a Berkeley Vision Project

UNIVERSITY of HOUSTON

UNIVERSIDAD NACIONAL DE COLOMBIA

HPC




```
# Parameters
learning_rate = 0.01
training_epochs = 25
batch_size = 100
display_step = 1
```

```
train_y_ohe = one_hot_encode_object_array(train_y)
test_y_ohe = one_hot_encode_object_array(test_y)

model = Sequential()
model.add(Dense(16, input_shape=(4,)))
model.add(Activation('sigmoid'))
model.add(Dense(3))
model.add(Activation('softmax'))
model.compile(loss='categorical_crossentropy', metrics=['accuracy'], optimizer='

# Actual modelling
model.fit(train_X, train_y_ohe, verbose=0, batch_size=1)
score, accuracy = model.evaluate(test_X, test_y_ohe, batch_size=16, verbose=0)
```

```
# Gradient Descent
optimizer = tf.train.GradientDescentOptimizer(learning_rate).minimize(cost)

# Initializing the variables
init = tf.initialize_all_variables()
```

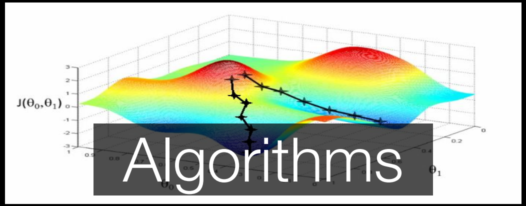
x
w
0 TensorType(float64, matrix)
1 TensorType(float64, vector)
dot
b
Elemwise{gt,no_inplace}
TensorType(int8, vector)

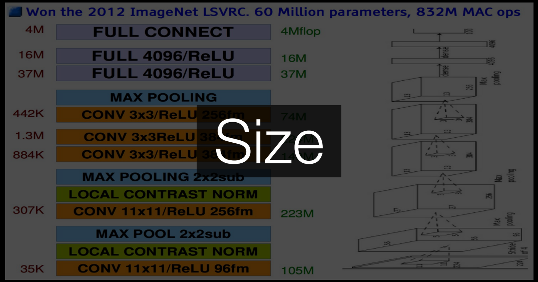
TensorFlow

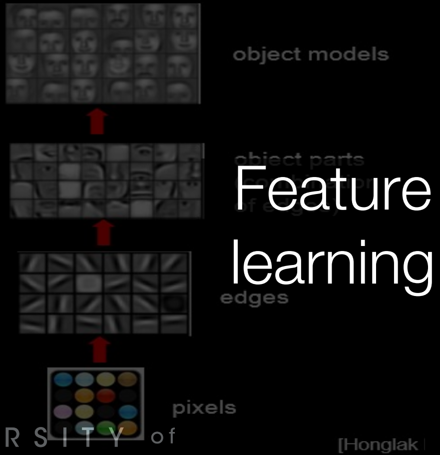
theano

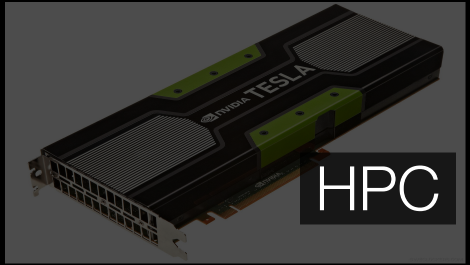
NVIDIA CUDA

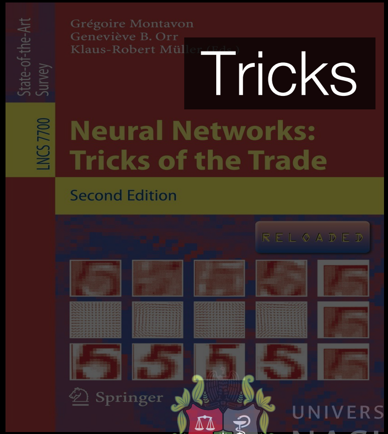
Deep learning recipe





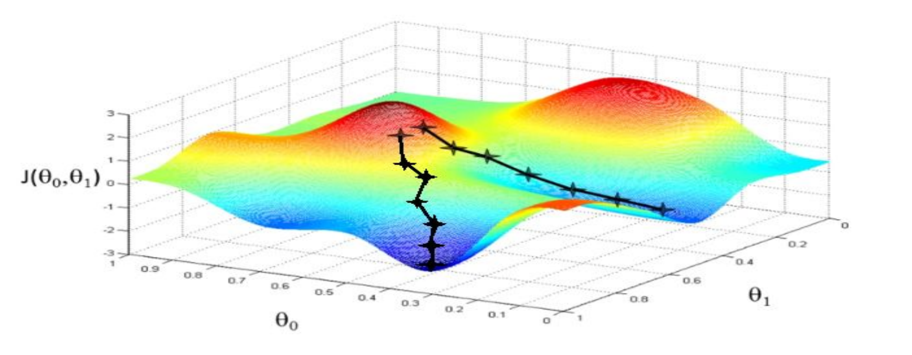




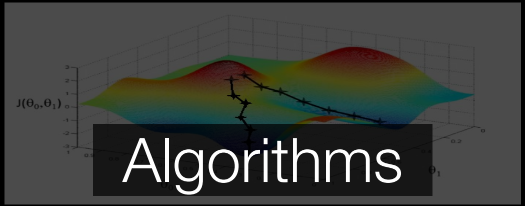


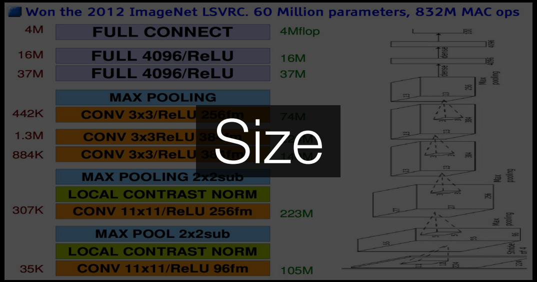
Algorithms

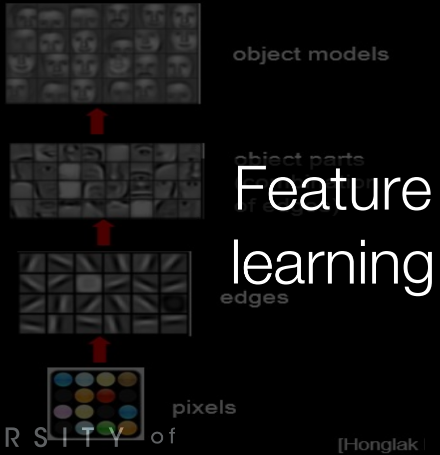
- Backpropagation
- Backpropagation through time
- Online learning (stochastic gradient descent)
- Softmax

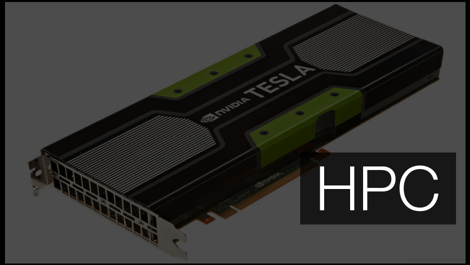


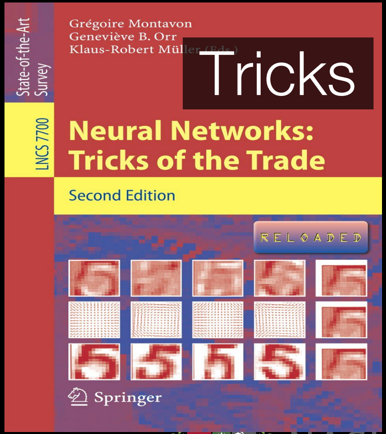
Deep learning recipe









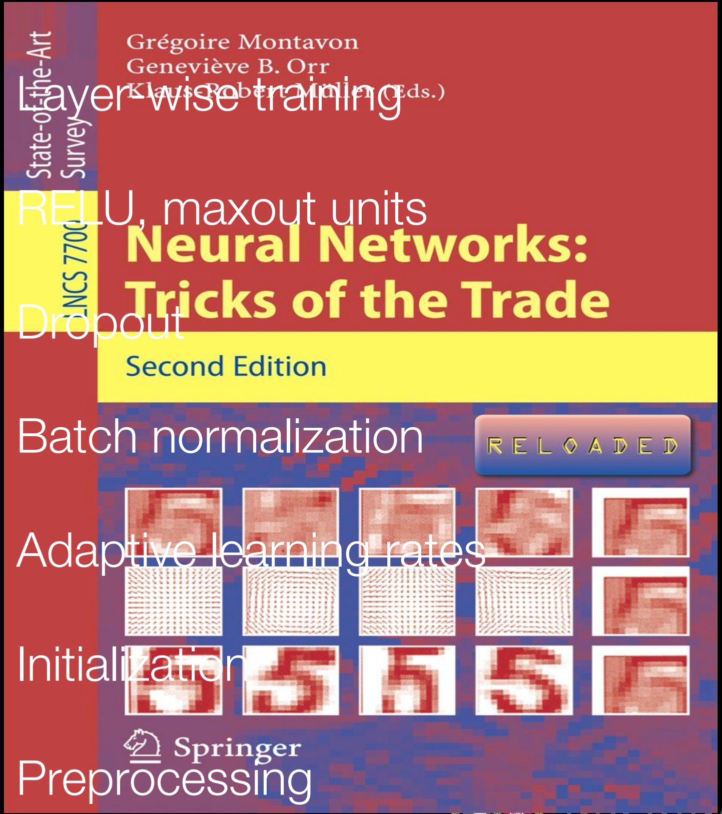


UNIVERSITY of HOUSTON

IDAD NACIONAL DE COLOMBIA

Tricks

- DL is mainly an engineering problem
- DL networks are hard to train
- Several tricks product of years of experience



- Layer-wise training
- RELU, maxout units
- Dropout
- Batch normalization
- Adaptive learning rates
- Initialization
- Preprocessing

Applications

- Computer vision:
  - Image: annotation, detection, segmentation, captioning
  - Video: object tracking, action recognition, segmentation
- Speech recognition and synthesis
- Text: language modeling, word/text representation, text classification, translation
- Biomedical image analysis

Natural Language Processing (NLP)

A quick but not so dirty intro

UNIVERSITY of HOUSTON

UNIVERSIDAD NACIONAL DE COLOMBIA

What is NLP?

- ★ Automated processing of human language (computational linguistics)
- ★ Computer science subfield that draws on knowledge from AI and Linguistics
- ★ Ultimate goal: To design programs that can take as input human language (any modality and language) and perform a useful task.

Why NLP?

- “ … language is what made us human” (Guy Deutscher)
- Through language humans:
  - Pass on knowledge
  - Create new thoughts and ideas
  - Express deep (and not so deep) reflections
- Practical value:
  - Companies want to know what consumers are saying
  - Intelligence communities want to know what persons of interest are planning
  - New products that use language as the interface with humans
- Scientific value:
  - Gain a deeper understanding of how the human brain is able to process language

Levels of Analysis

- ★ Speech
  - Phonology
- ★ Text
  - Morphology: the structure of words
  - Syntax: how these sequences are structured
  - Semantics: meaning of the strings
  - Pragmatics: discourse
  - Interaction between levels

Some NLP applications

- Speech recognition
  - Voicemail transcription
- Dialogue systems
  - Siri, Cortana
- Information extraction
  - Named Entity Recognition and Linking
  - Event detection
- Machine translation
  - Text to text
  - Speech to speech

Challenges in NLP

Main issue is ambiguity ….

Ambiguity in Speech

- ★ 264 Lane Street vs. 26 four-lane street
- ★ For invoices vs foreign voices
- ★ Colorectal cancer risks vs co-director cancel risks
- ★ Frapuccino vs Fred Paccino

Ambiguity in Morphological Analysis

ride	rideable
do	doable
like	likeable

- Pattern: Verb + “able” → Adjective (able to do/be Verb-ed)

Ambiguity in Morphological Analysis

happy	unhappy
cool	uncool
stable	unstable

- Pattern: “un” + Adjective → Adjective (not Adjective)

Ambiguity in Morphological Analysis

do	undo
zip	unzip
dress	undress

- Pattern: “un” + Verb → Adjective (to reverse Verb-ing)

Ambiguity in Morphological Analysis

What about the word unlockable?

Image source: http://plywoodchair.com/wp-content/uploads/2015/03/Garage-Door-Lock-Mechanism-Mesmerizing-Door-Design-Ideas.jpg

Ambiguity in Morphological Analysis

Option 1:

“un” + lock (Verb) → unlock (Verb) (to reverse locking)

unlock + “able” → unlockabe (Adjective) (able to unlock)

Option 2:

lock + “able” → lockable (Adjective) (able to lock)

“un” + lockable → unlockable (Adjective) (not able to lock)

Ambiguity in Syntax

Ambiguity in Syntax

- ★ Jake told Mike he has cancer
- ★ Eat spaghetti with meatballs vs eat spaghetti with chopsticks
- ★ We saw the Eiffel Tower flying to Paris
- ★ Old men and women

Interesting Advances in Deep Learning for NLP

Sentiment analysis:

http://nlp.stanford.edu:8080/sentiment/rntnDemo.html

Poetry generation:

http://52.24.230.241/poem/

Interesting Advances in Deep Learning for NLP

★ Lowest reported WER in speech recognition (5.1)
  - Neural network acoustic and language models

Microsoft researchers achieve new conversational speech recognition milestone

August 20, 2017 | Posted by Microsoft Research Blog

Share Tweet Share

By Xuedong Huang, Technical Fellow, Microsoft

Last year, Microsoft's speech and dialog research group announced a milestone in reaching human parity on the Switchboard conversational speech recognition task, meaning we had created technology that recognized words in a conversation as well as professional human transcribers.

After our transcription system reached the 5.9 percent word error rate that we had measured for humans, other researchers conducted their own study, employing a more involved multi-transcriber process, which yielded a 5.1 human parity word error rate. This was consistent with prior research that showed that humans achieve higher levels of agreement on the

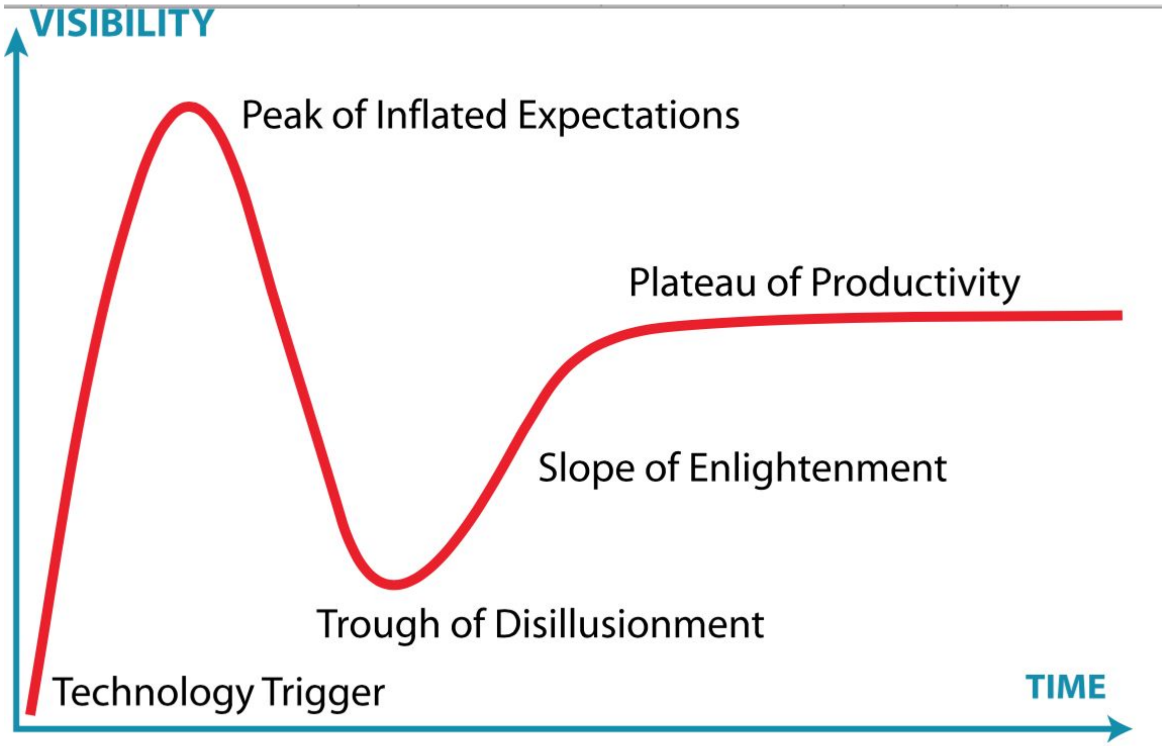
Interesting Advances in Deep Learning for NLP

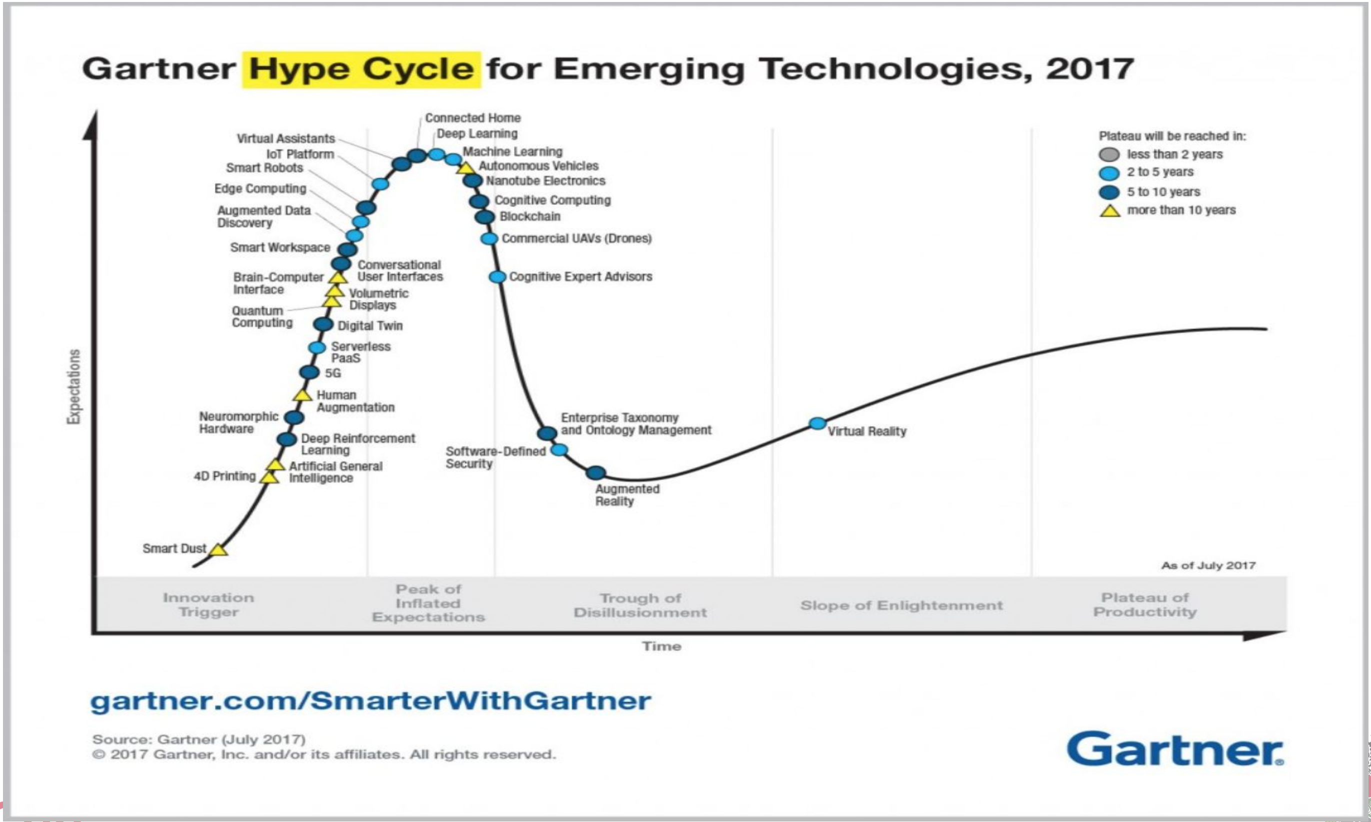
- ★ Series of first places on different shared tasks:
  - Sentiment Analysis on Financial Data (SemEval-2017 Task 5)
  - Novel and Emerging Named Entity Recognition (2017 WNUT at EMNLP'17)
  - Sentiment Analysis on Twitter (SemEval-2017 Task 4)

Deep Learning Hype

- ★ Most papers in NLP related conferences use DL
- ★ Plenty of job opportunities:
  - http://deeplearning.net/deep-learning-job-listings/
  - Linkedin shows > 2,500 matches

Technology Hype





UNIVERSITY of HOUSTON

UNIVERSIDAD NACIONAL DE COLOMBIA

Cautionary Tale

- ★ Misinformed predictions about DL
- ★ Interpretability
- ★ Domain knowledge:
  - An Adversarial Review of “Adversarial Generation of Natural Language”
    - The paper oversells what they accomplish
    - Lack of understanding of the domain → failure to devise proper evaluation
    - Sexy models should not be the goal
    - Review process is “damaged” by loss of anonymity in arxiv

Course Administrivia

Course Info

- Website: https://fagonzalezo.github.io/dl-tau-2017-2/
- Structure and Grading
  - 3 assignments: 45% (15% each)
  - One mid term exam: 20%
  - Paper presentation: 10 %
  - End of semester project: 25% (includes final report and poster)

Assignments

★ Assignment 0 is a warm up exercise
★ Assignment 1 (NN basics, word embeddings and text classification)
★ Assignment 2 (language modeling and generation)
★ Assignment 3 (semantic similarity)

Note that assignments up to 1 day late will receive up to 80% of the credit, and 0 credit after 1 day late.

Paper Presentations

Choose a paper to present to the class in <10 minutes. The list of possible papers to choose from will be posted in the course website.

Each student will need to present one paper.

Final Project

★ Individual Projects on a research topic chosen in discussion with the instructors.

★ Students need to submit a proposal due Nov. 10th. Final project poster presentations, report and github repository are Dec. 11th.