Deep learning for NLP

Eneko Agirre, Gorka Azkune
Ander Barrena, Oier Lopez de Lacalle

@eagirre @gazkune @oierldl #dl4nlp

http://ixa2.si.ehu.eus/eneko/dl4nlp

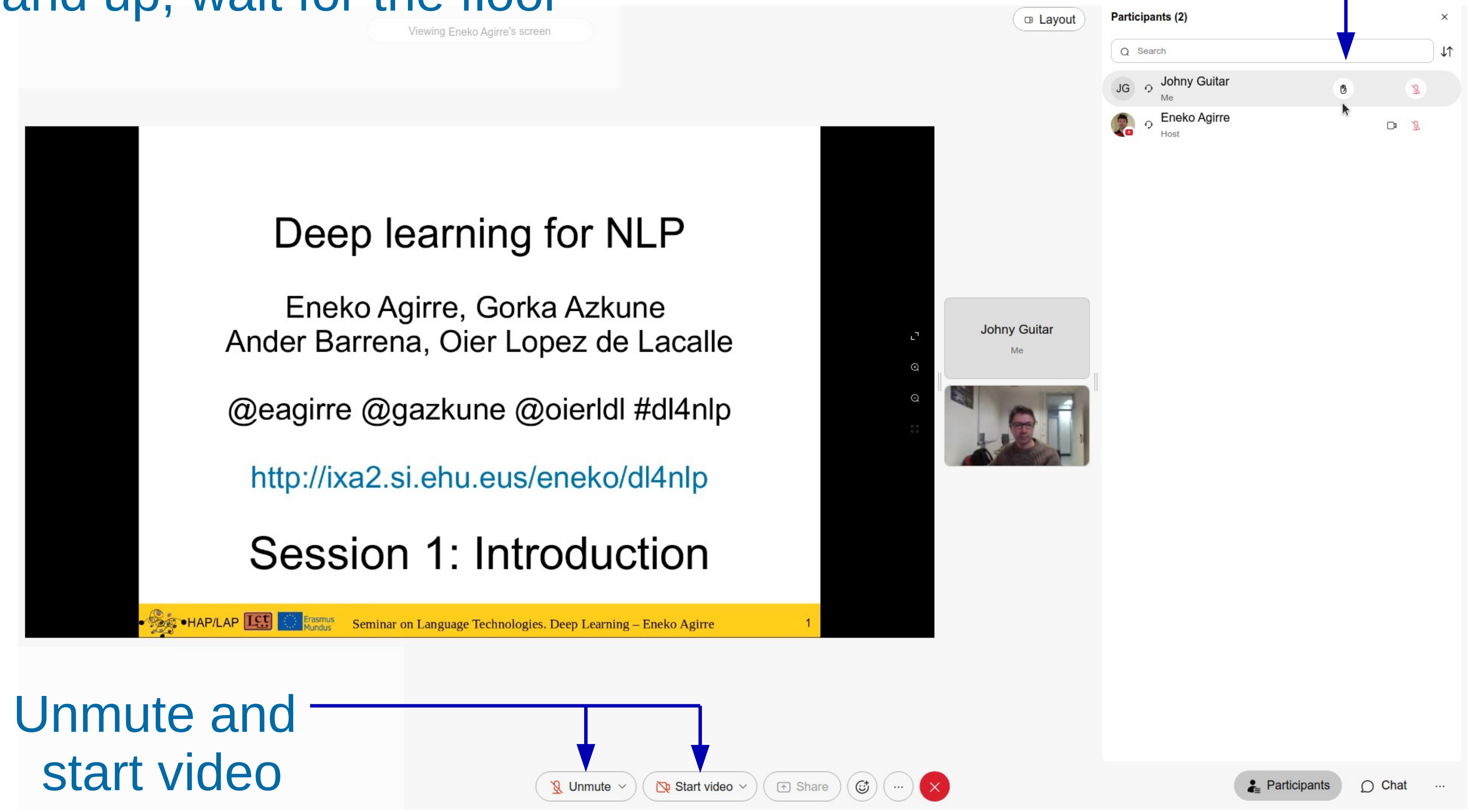
Session 1: Introduction

Practicalities: class protocol

- We have a full house! 35 master students, 25 registered attendants plus some guests from our research center
- Those onsite, please minimize noise (computers off)
- Those online, please keep the mic and camera off
- You are encouraged to interrupt, make questions and comments. Anytime!
  - Please use the "raise hand" icon (participants window), and wait until the instructor gives you permission.
  - The instructor will then give you the floor: enable your mic and camera, so the rest can hear and see you.

Practicalities: class protocol

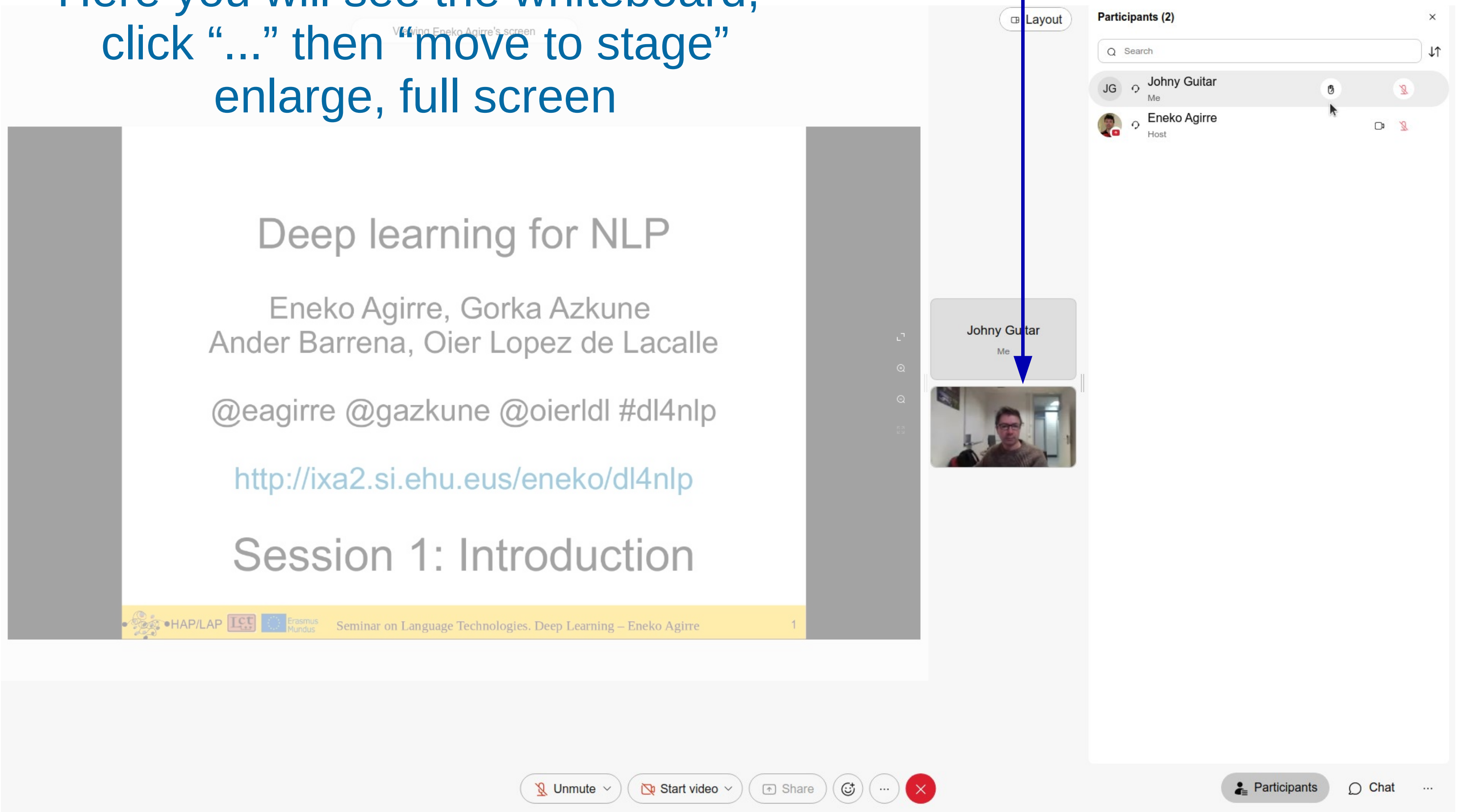
Hand up, wait for the floor



Unmute and start video

Practicalities: class protocol

Here you will see the whiteboard,
click "..." then "move to stage"
enlarge, full screen



Practicalities: attendance control

- We need that you identify when entering with your SURNAME
  - If not, close the window now and open a new one
- Otherwise we won't be able to provide attendance certificate
- Because link is open to anyone, we might need to remove SURNAMES not in the list

Plan for the course

- Introductory crash course
  on deep learning for natural language processing
- Allow to understand
  the latest developments in deep learning
  - Not only use pre-existing neural networks
    but be able to reimplement and adapt
- Provide leads to explore and learn further
  - Master projects ideas welcome!
  - Open for collaborations hitz.eus

Plan for the course

- 7 theoretical sessions (Eneko, Gorka) (approx. 150 min, time for questions)
- 7 hands-on laboratories (150 minutes)
  - Master students => Onsite (Oier)
  - Independent course => Online (Ander)

- Calendar with slides and labs:: http://ixa2.si.ehu.eus/eneko/dl4nlp
- Note on downloading/uploading the labs data folder => do it before lab, at home

Labs and pre-requisites

- Basic programming experience, university-level course in computer science, experience in Python. Basic math skills (algebra or pre-calculus)
- Laboratories:
  - Python and Tensorflow, using servers from Google Colaboratory
  - We start from easy to difficult
  - Time might be tight => auto-study / finish labs at home
  - Time might be plenty => review slides / do assignments

Evaluation

Depends on student profile:

- **Independent course:** Attendance or Progress certificate
  - Attendance => attendance certificate
  - 7 labs in class => progress cert. **Deadline:** prior to next lab
- **Master students**: grades of labs, assignments and project
  - 7 labs in class => 4 points. **Deadline:** prior to next lab
  - 4 assignments => 2 points. **Deadlines** for full grade:
    1&2 by end of Jan., 3&4 by end of Feb.
    No submission beyond 26th of March
  - Personal work: compulsory
    Presentations: 17th and 18th of March (further communication)

Certificates (independent course)

- University releases formal certificates for a fee
- Requires us to use signature sheet
- Administrative regulations cause delays, they will be ready beginning of March

http://www.ehu.eus/eu/web/complementarios/ziurtagiri-eskaria

(check English option left-top corner)

Quiz

- How many of you have done
  - a course on deep learning
  - a course on natural language processing

Quiz

- How many of you have done
  - a course on deep learning
  - a course on natural language processing

  - a personal implementation project with Python
  - a personal implementation project on deep learning

Plan for the course

- Introduction: machine learning and NLP
- Multilayer perceptron
- Word representation and
  Recurrent neural networks (RNN)
- Sequence-to-Sequence (seq2seq) and
  Machine Translation
- Attention, transformers and
  Natural language inference
- Pre-trained transformers, BERT, GPT
- Bridging the gap between natural languages
  and the visual world

Session 1
Introduction: Machine Learning for NLP

http://ixa2.si.ehu.eus/eneko/dl4nlp

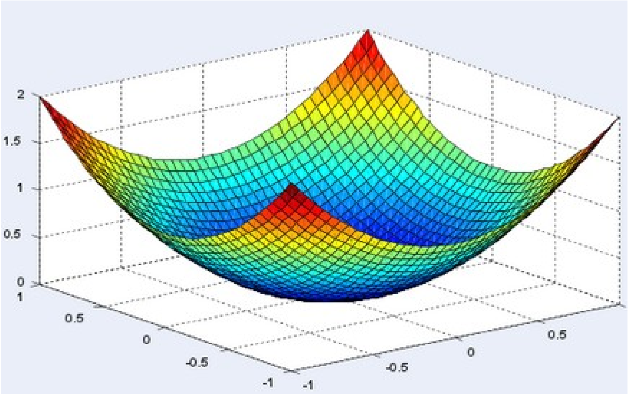
Introduction – plan for this session

- Machine learning, Deep learning
- Natural Language Processing
- A sample NLP task with ML
  - sentiment analysis
  - features (bag of words)
  - classification with logistic regression
- Tensorflow (tomorrow)

What is deep learning

A subfield of machine learning

- Supervised ML, given a dataset of examples **x** with labels **y**
- Learn a function **f(x)**→**y** with low training error and low test error

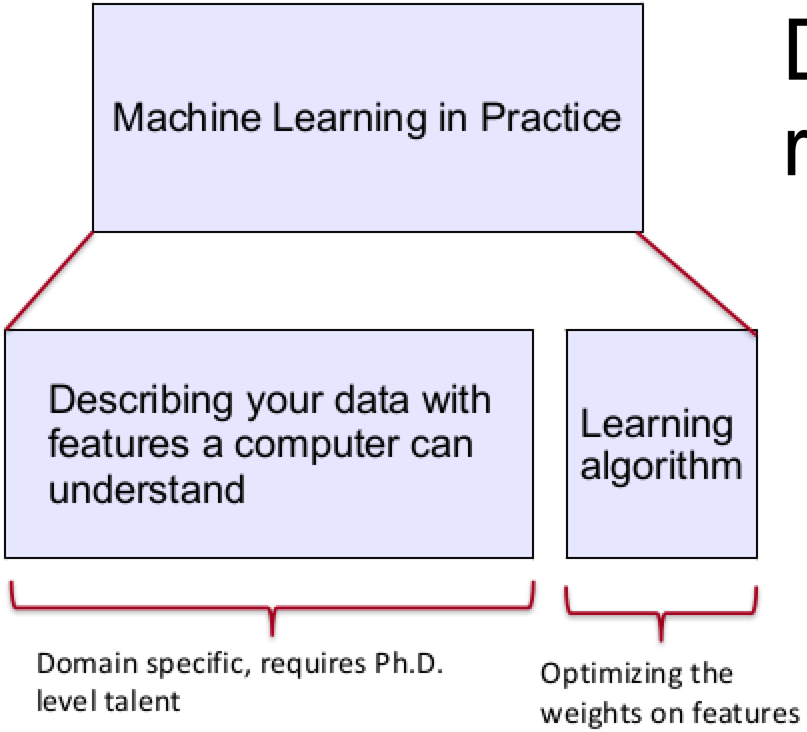


Source: staesthetic.wordpress.com

Key manual step: design features to extract key information from **x** (**representation**)

- e.g. weather forecast (wind, temperature, humidity, pressure, precipitations … – local and nearby locations)
- e.g. sentiment of tweet (keywords like “good” “bad”, certain emojis, ...)

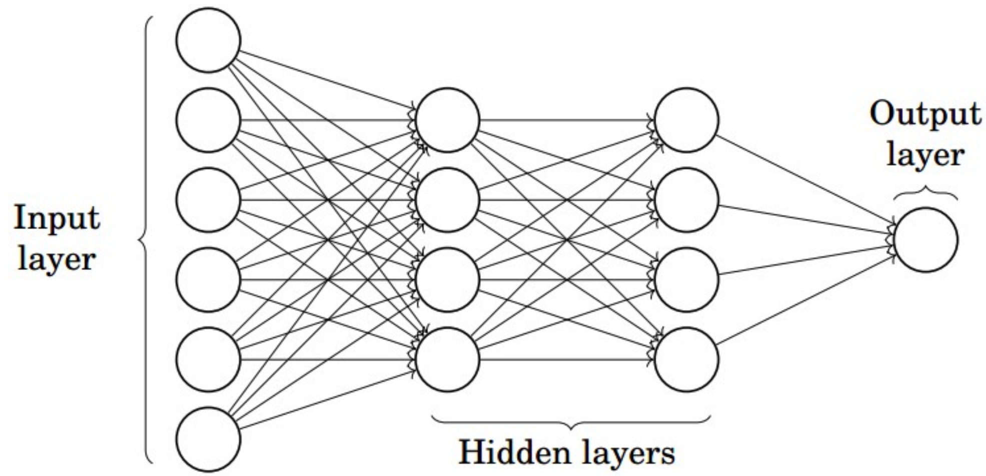
What is deep learning



Source: Chris Manning cs224n

Deep learning jointly learns representation and output

x **f(x)=y**



Source: www.vaetas.cz

Deep? Multiple levels

What is deep learning

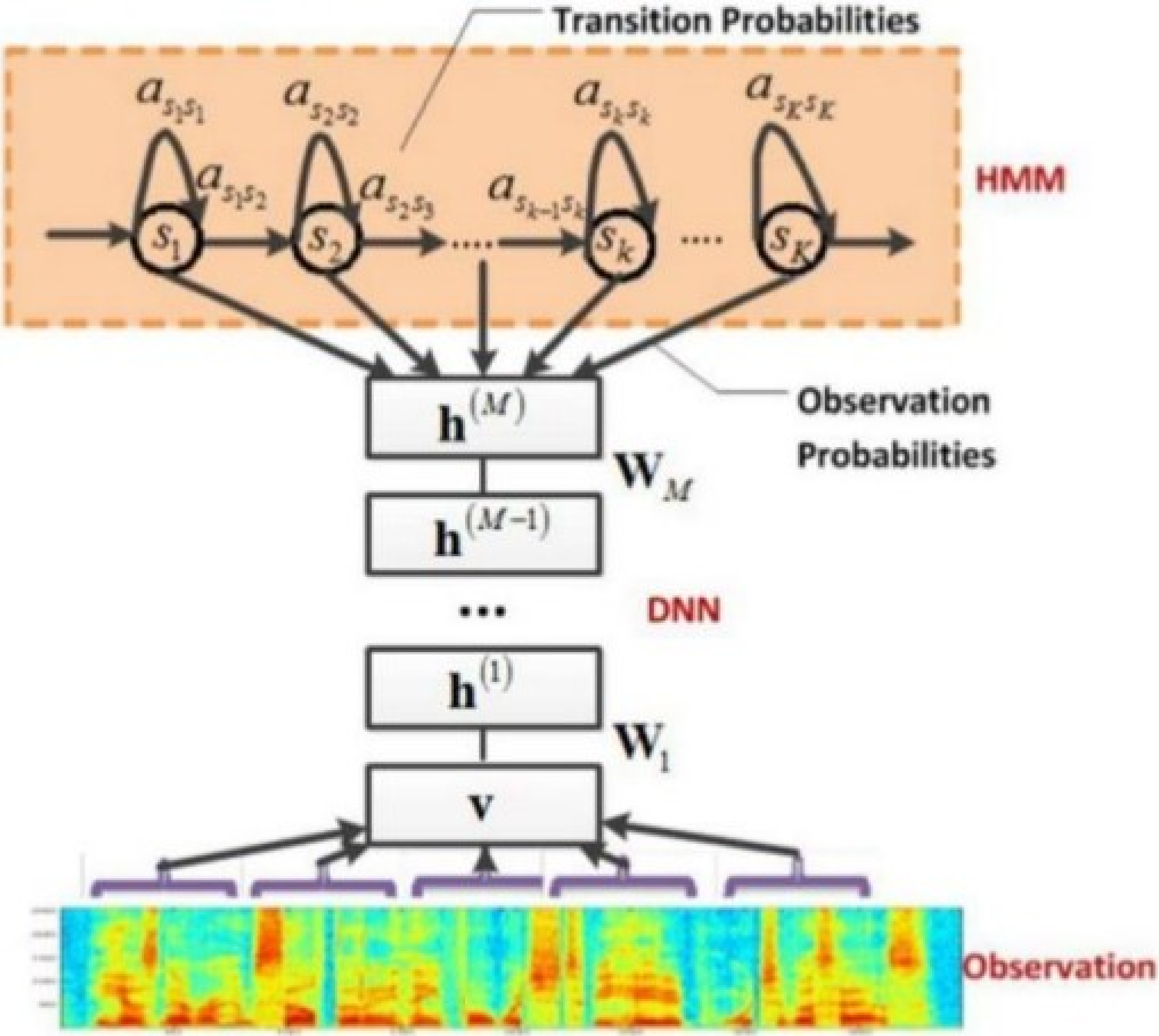
- A rebranding of artificial neural networks with more than two layers
  - Differentiable
  - Parameters by stochastic gradient descent
- Might be rebranded as Differentiable Programming
  - Combine parameterized functional blocks
  - Structure dependent of input (Dynamic networks)
  - Imperative differentiable programming languages

(source: Yann LeCun 5/1/2018 via facebook)

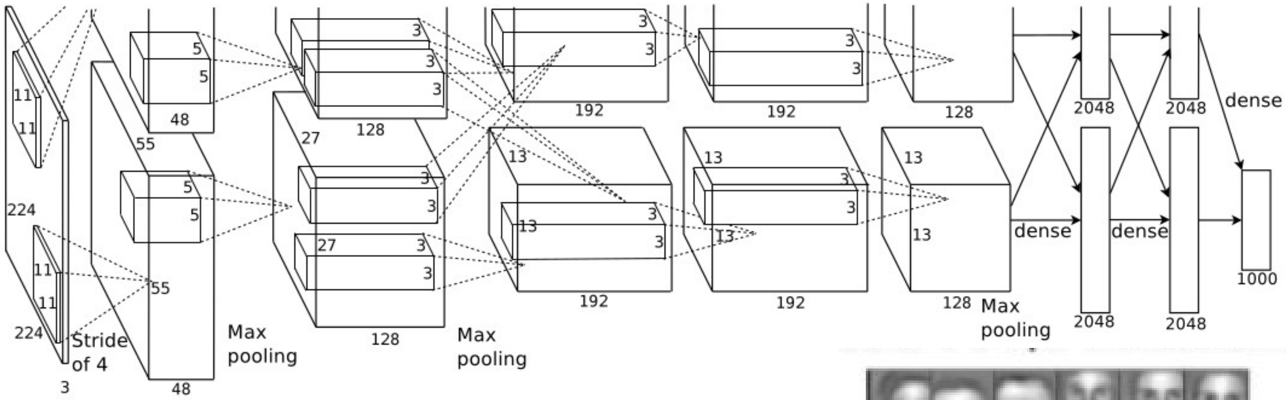
What is deep learning

Why now?

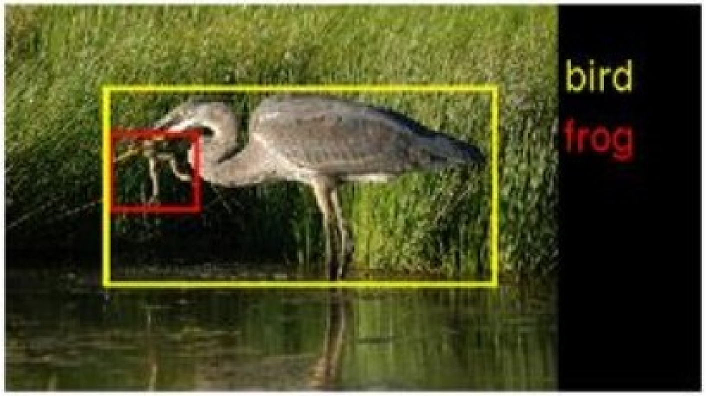
- Large amounts of data
- Multicore CPUs and GPUs
- Best performance: speech, vision, NLP

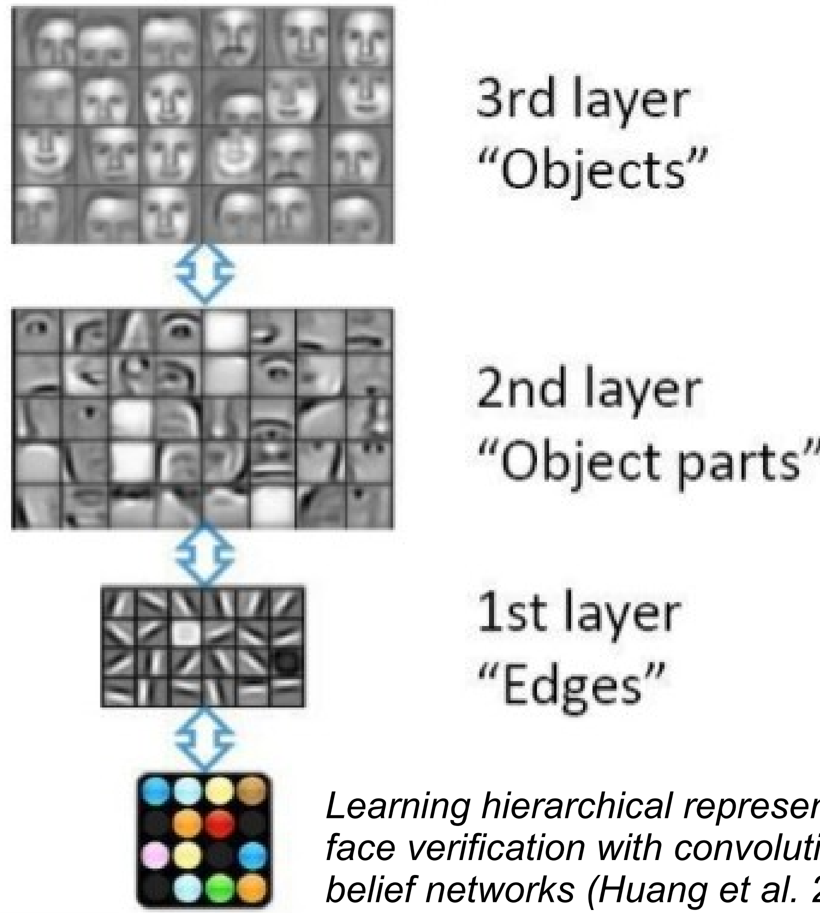


Dahl, George E., et al. "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition." *Audio, Speech, and Language Processing, IEEE Transactions on* 20.1 (2012): 30-42.



ImageNet Classification with Deep Convolutional Neural Networks (Krizhevsky et al. 2012)





Learning hierarchical representations for face verification with convolutional deep belief networks (Huang et al. 2012)

What is deep learning

Deep learning is behind the latest Artificial Intelligence hype

Connectivity

Amazon's cashier-less Seattle grocery store is opening to the public

At Amazon Go, you grab your milk and leave. It might take some getting used to.

by Rachel Metz January 21, 2018

Source: www.technologyreview.com

What is deep learning

Deep learning is behind the latest Artificial Intelligence hype

Elon Musk, founder, CEO and lead designer at SpaceX and co-founder of Tesla, speaks at the International Space Station Research and Development Conference in Washington, U.S., July 19, 2017 / *REUTERS/Aaron P. Bernstein*

'Should that be controlled by a few people at Google with no oversight?'

AATIF SULLEYMAN
Friday 24 November 2017 19:01 GMT

3K SHARES Atsegin CLICK TO FOLLOW THE INDEPENDENT TECH

Source: www.independent.co.uk

What is deep learning

Deep learning is behind the latest Artificial Intelligence hype

6,045 views | Jun 25, 2018, 02:10am

The AI Skills Crisis And How To Close The Gap

Bernard Marr Contributor

Now that nearly every company is considering how artificial intelligence (AI) applications can positively impact their businesses, they are on the hunt for professionals to help them make their vision a reality. According to research done by Glassdoor, data scientists have the No. 1 job in the United States. The survey looked at salary, job

Source: www.forbes.com

Introduction – plan for this session

- Machine learning, Deep learning
- **Natural Language Processing**
- A sample NLP task with ML
  - sentiment analysis
  - features (bag of words)
  - classification with logistic regression
- Tensorflow (tomorrow)

Natural Language Processing

- Intersection of AI and linguistics
  - Machine learning
  - Big data
  - Data science

  aka text processing, text analytics,
  natural language engineering,
  computational linguistics, …

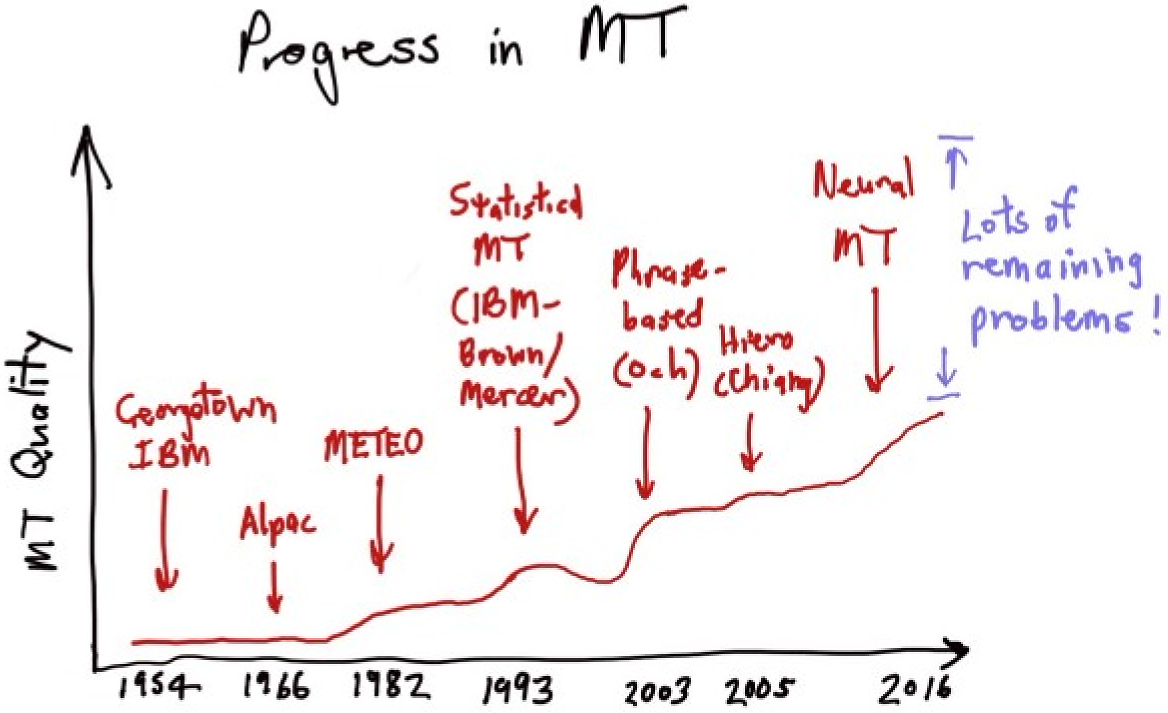
- Goal: process natural language input
  to perform a task

- Language understanding is AI-complete

  Youth unemployment around 27%, wow!

NLP tasks – Machine Translation

Given sentence or sets of sentences output translation in other languages

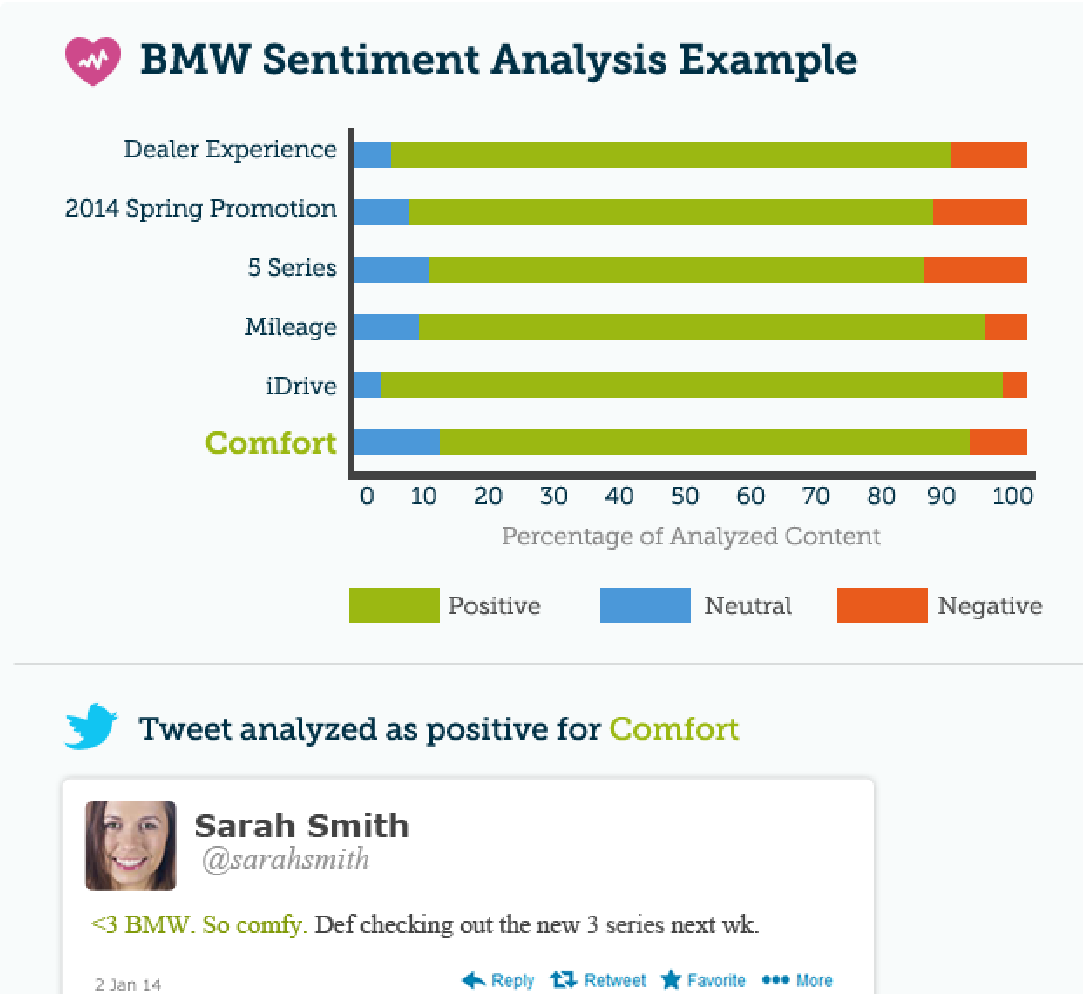


Source: Chris Manning

NLP tasks – Sentiment Analysis

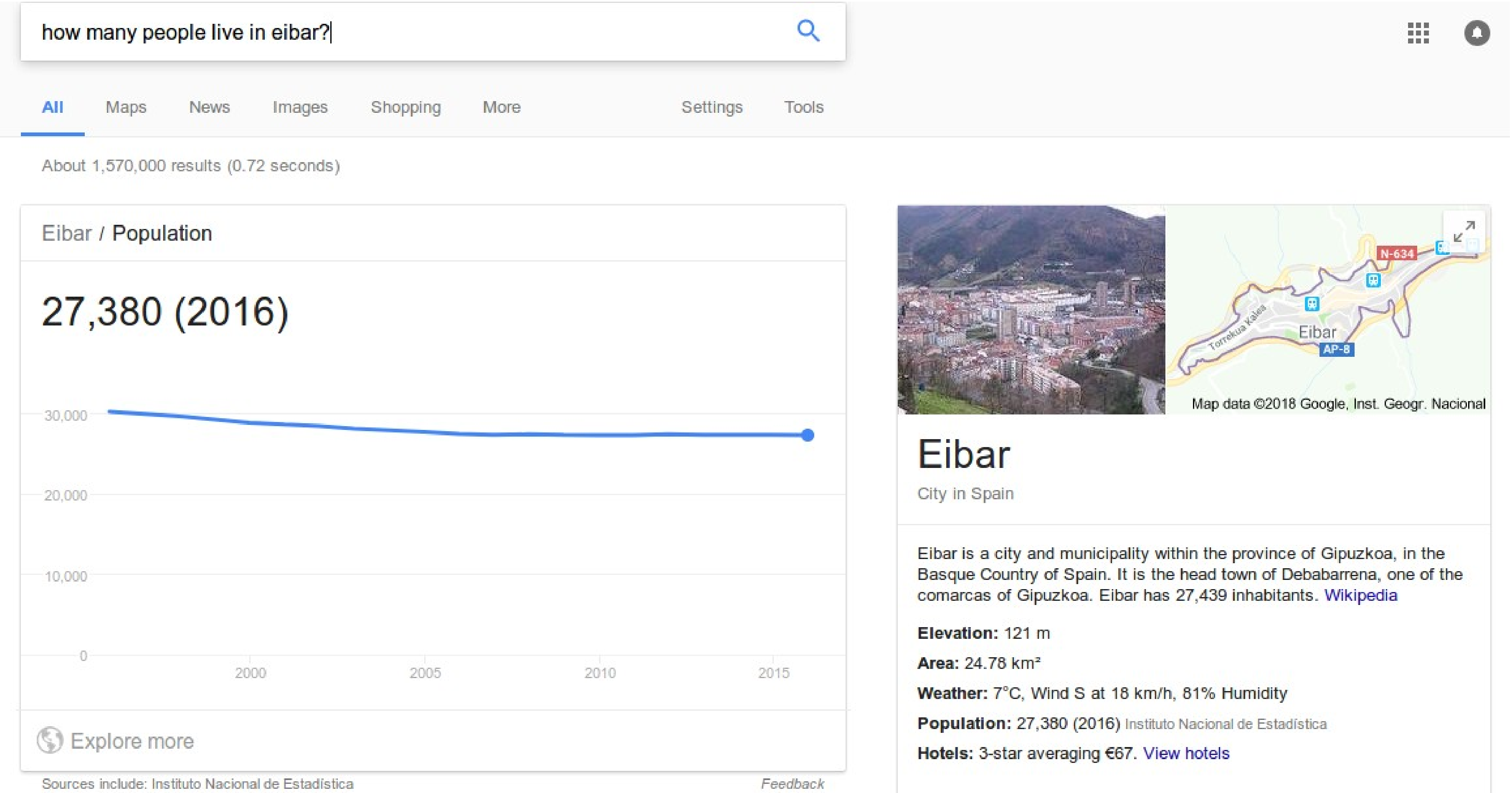
Given sentence or short document

text is positive/ negative/neutral



Source: www.crowdsource.com

NLP tasks – Question Answering



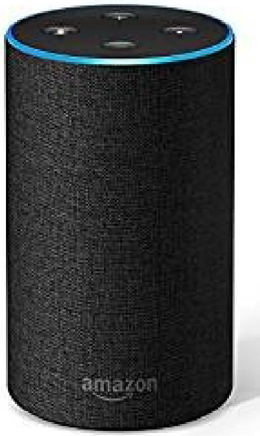
Source: google.com

NLP tasks – (Spoken) dialogue

Simple dialogues with Alexa, OK google, Siri:

What's in the news?
What's the weather like?
What's my commute look like?
Add eggs to my shopping list
I need to buy laundry detergent.

Source: www.cnet.com/how-to/
the-complete-list-of-siri-commands/
complete-list-of-ok-google-commands/
amazon-echo-the-complete-list-of-alexa-commands/



Alexa Prize: socialbots

User: Let's chat about the Mars Mission.

Socialbot: There are multiple Mars missions, some public and some private.

User: Who do you think will succeed?

Socialbot: I think more than one will succeed. The first one to get to Mars will open the doors for others.

User: I'd love to go to Mars.

Socialbot: Me too. Luckily I can hitch a ride on any computer.

User: That's very funny.

Socialbot: The biggest challenge isn't technology, it's funding. It will cost $500 billion to send humans to Mars.

NLP tasks – getting mainstream

- Search engines
- Machine (assisted) translation
- Online advertisement placement
- Sentiment analysis for finance, reputation
- Speech recognition, dialogue

NLP levels

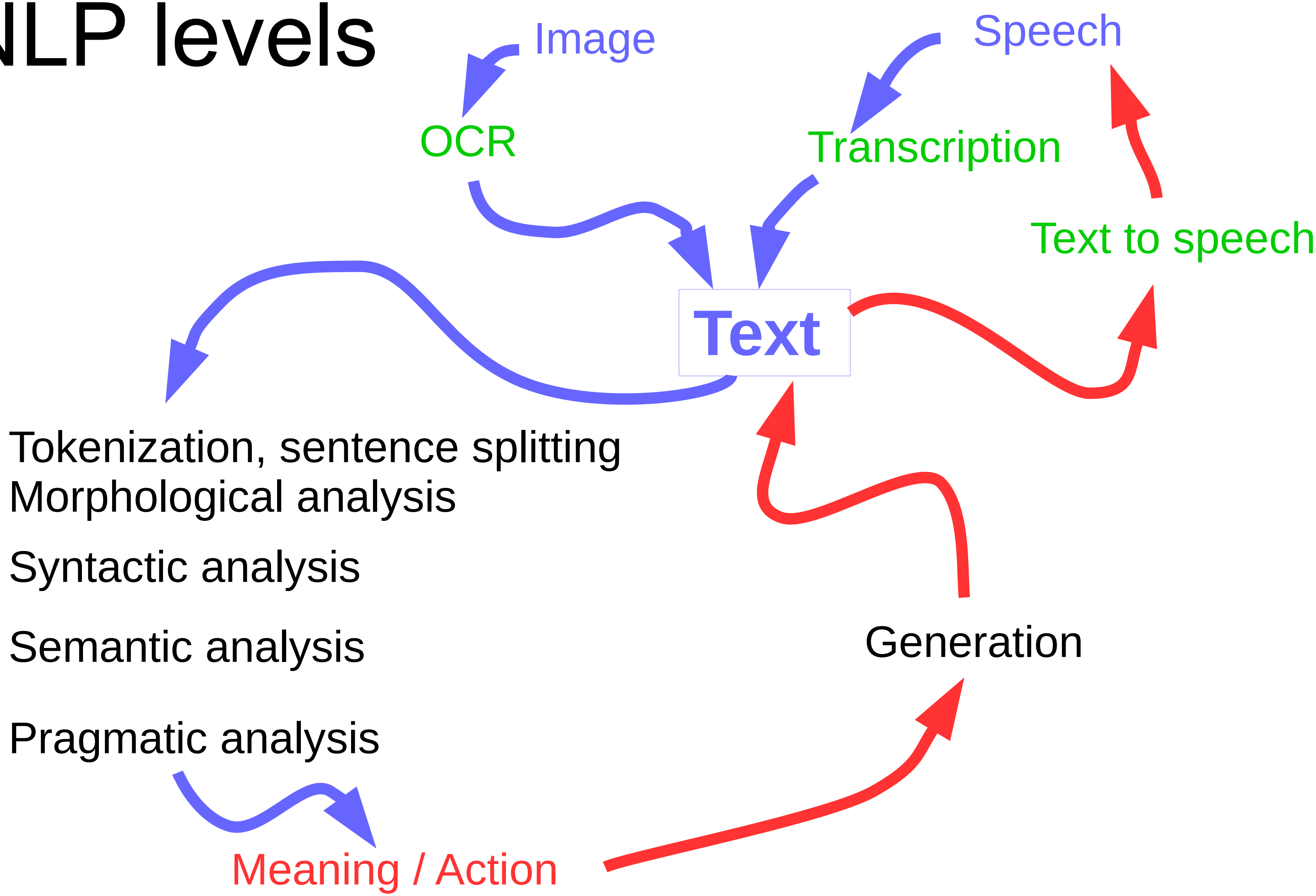


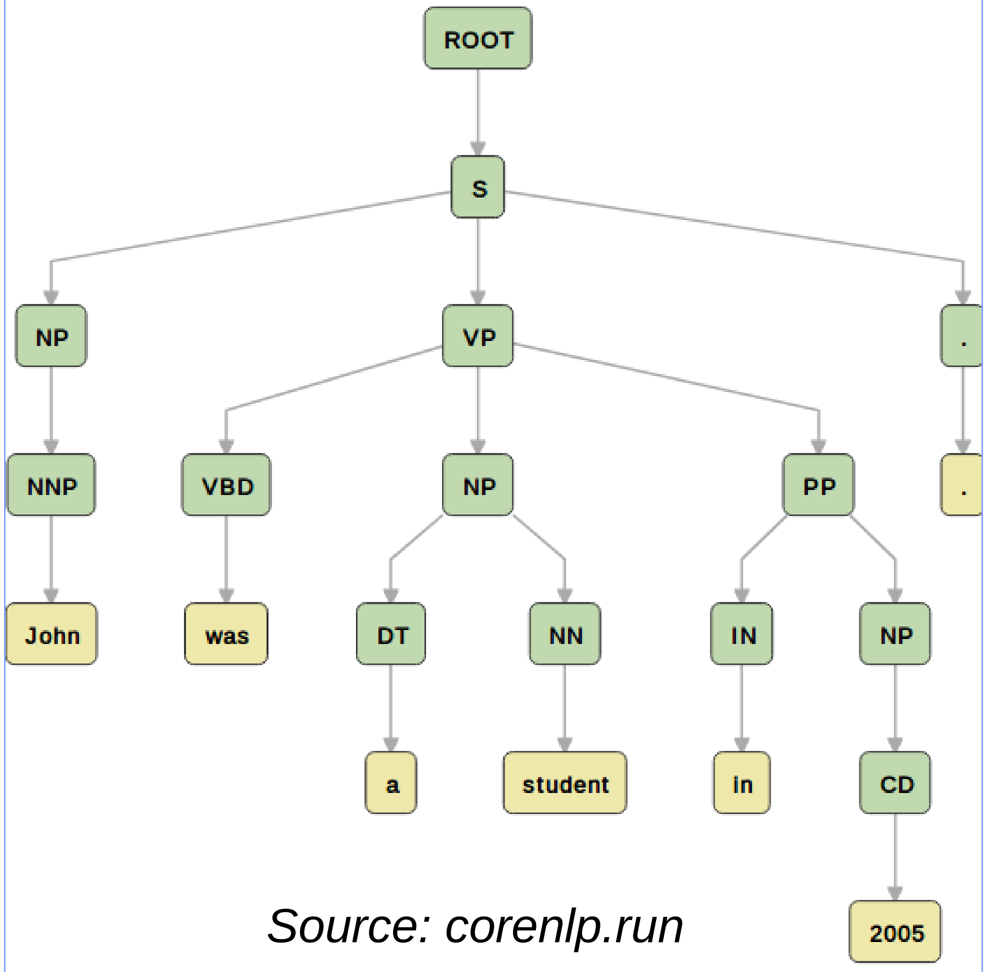
Image
Speech
OCR
Transcription
Text to speech
Text
Tokenization, sentence splitting
Morphological analysis
Syntactic analysis
Semantic analysis
Pragmatic analysis
Generation
Meaning / Action

NLP levels

- Tokenization, sentence splitting
  - *John was a student in 2005.*
  - *2005. urtean John ikaslea zen.*
  - 约翰是 2005 年的学生。
- Morphological analysis (PoS, lemma, NERC)
  - *was* (lemma *be,* PoS *V)*
  - *John F. Kennedy* (lemma *John_F._Kennedy*, PoS *NNP*, NERC *Person*)
  - *etxekoarentzat* (lemma *etxe*, PoS N, for the one of the house)
  - *tomemos* (lemma *tomar*, PoS V, let we take it)

NLP levels

- Syntactic analysis



Source: corenlp.run

- Semantic analysis

$$\exists x, y \wedge name(x, \mathrm{John}) \wedge student(x)$$
$$intime(x, y) \wedge time(y, \mathrm{T2005xxxx})$$

Source: gmb.let.rug.nl

NLP levels

- Pragmatics (discourse, correference, ...)
  - ***But** Mary become a lawyer **that year***.
  - Wasn’t it one year later?
- Inference
  - *John and Mary were law students in Dec. 2005*
  - *Mary was working full-time as a lawyer in 2005*

NLP is difficult

- Ambiguity at all levels
  - *cells in prisons* vs. *cells in animals*
  - *One morning I shot an elephant in my pajamas.*
    How he got into my pajamas I'll never know. (Groucho)
  - *You mean Mary Smith or Mary Doe?*
- Variability at all levels (many ways to convey a meaning)
- Subtlety
- Understanding language requires
  - Language knowledge (word meaning, grammar, ...)
  - World knowledge (physical, encyclopedic, visual ...)
  - Common sense and inference ability
- But sometimes it is surprisingly easy!

Brief history of NLP

- 1960s: Complex rules and first order logic.
  Humans build complex grammars.
- 1990s: Supervised machine learning.
  Humans annotate text,
    design laborious task-specific features,
    and apply ML techniques.
- 2010s: Deep learning.
  Learning continuous representations,
  get rid of task-specific features.

Why deep learning for NLP?

- Technology behind current
  speech processing and machine translation
  - Advances in the state-of-the-art of most tasks
- Focus on representation learning
  - Learns a representation for words and word sequences
    that is fitted to the task
    … including world and visual knowledge.
  - Naturally accounts for graded judgements about language:
    - Word similarity: building / house
    - Sentence similarity:
      A pony is close to the house / There is a horse in the front yard
- End-to-end joint learning (vs. pipeline)
- Transfer models between tasks (word embeddings)
- But deep learning has its limitations too!!

Quiz

Mention one NLP task of your interest

Introduction – plan for this session

- Machine learning, Deep learning
- Natural Language Processing
- **A sample NLP task with ML**
  - sentiment analysis
  - features (bag of words)
  - classification with logistic regression
- Tensorflow (tomorrow)

A sample NLP task with ML

Text classification

- Spam or not
- Positive or negative movie review
  - Full of zany characters and richly applied satire
  - It was pathetic
  - Modestly accomplished, lifted by two terrific performances.
  - Not NEARLY as funny as its title

A sample NLP task with ML

Text classification:

- Input
  - A document d $\in$ D
  - A fixed set of classes $C=\{c_1, c_2, \ldots c_j\}$
- Output: a predicted class y $\in$ C

Text classification as regression:

- Input
  - A document d $\in$ D
  - A range of real values C=[0,j]
- Output: a predicted value y $\in$ C

A sample NLP task with ML

Text classification method 1: Hand-coded rules

- Rules based on combination of words and other features
  - emoji ∈ { 🙂 😂 } => positive
  - word ∈ { terrible, ugly } => negative
- Accuracy very high
  - If rules refined by expert
- But writing and maintenance very expensive

A sample NLP task with ML

Text classification method 2: Supervised ML

Learn a classifier from hand-annotated examples

- Input: A training set of n hand-labeled documents $(x_1,y_1) \dots (x_m, y_m)$
- Output: A learned classifier $f{:}D \rightarrow C$
- Recall very high
- Cost-effective: experts only needed for annotation

Supervised doc. classification

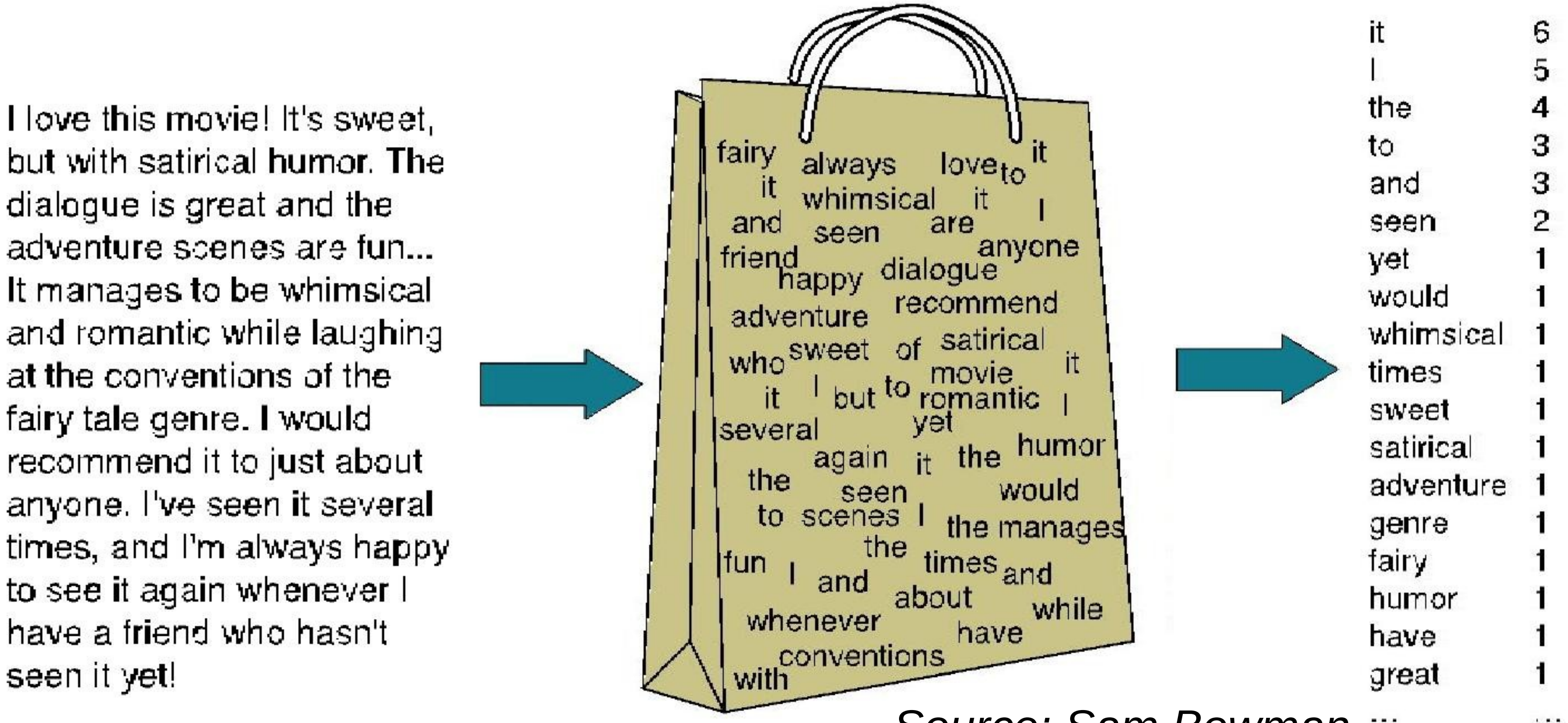
Representation of each example (document)

- Key idea for most machine learning
- Example as a vector of features x
  - All example same number of features
  - Features: boolean, integer, real
- Pre-processing code
  to convert from example into feature vector
  - Substantial effort

Supervised doc. classification

Representation of each example (document)

⇒ Bag of words representation



Source: Sam Bowman

Supervised doc. classification

Representation of each example (document)

⇒ Bag of words representation

Limited number of words (size of vocabulary)!

Each word is a feature:

non-negative integer or boolean:

it	6
I	5
the	4
to	4
and	3
...	

it	1	house	0
I	1	cat	0
the	1	mat	0
to	1	him	0
and	1	eibar	0
...		...	

Supervised doc. classification

Output of the classifier: continuous or discrete

Sentiment analysis:

real number [0:10] or two classes

it	1
I	1
the	1
to	1
and	1
...	
house	0
cat	0
mat	0
eibar	0
...	

f(...) = c

8.1

Positive!

Supervised doc. classification: continuous

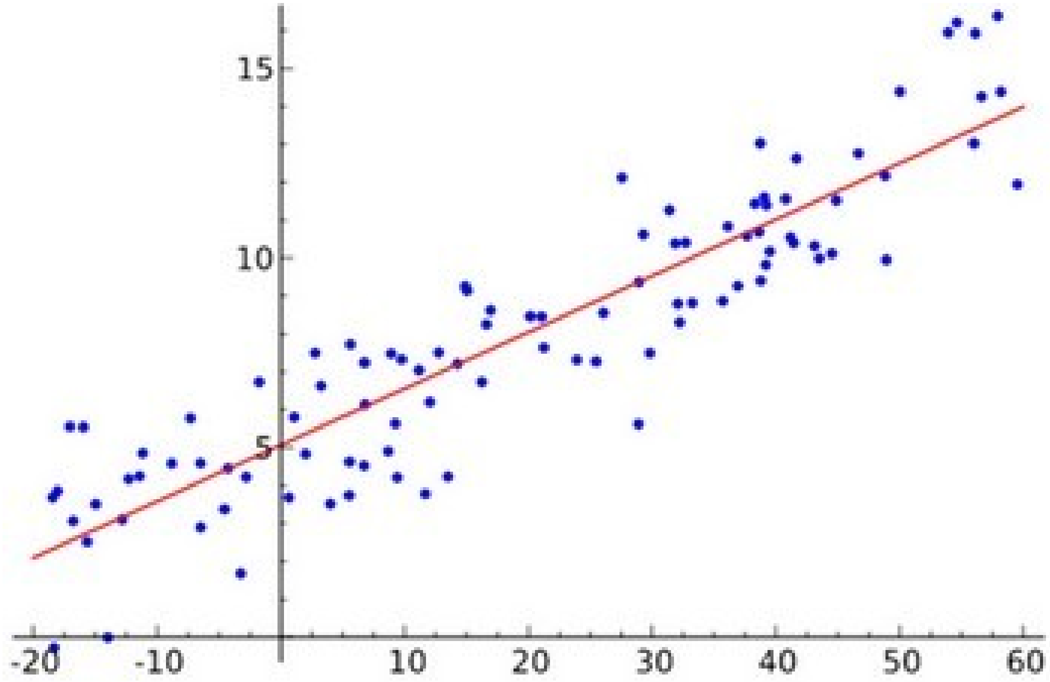
Linear regression (continuous class)

- Assign a weight to each feature: vector of weights w
- Output value dot product: $y = w^T x + b$
  - e.g. with three features: $y = w_1x_1 + w_2x_2 + w_3x_3 + b$
- Task for linear regression:
  - Given labeled examples
  - Find w such that the output value (y) is not too wrong for as many training examples as possible

Supervised doc. classification: continuous

Linear regression

- with one feature: $y = w_1x_1 + b$



Source: gerardnico.com/wiki/data_mining/linear_regression

Supervised doc. Classification: discrete

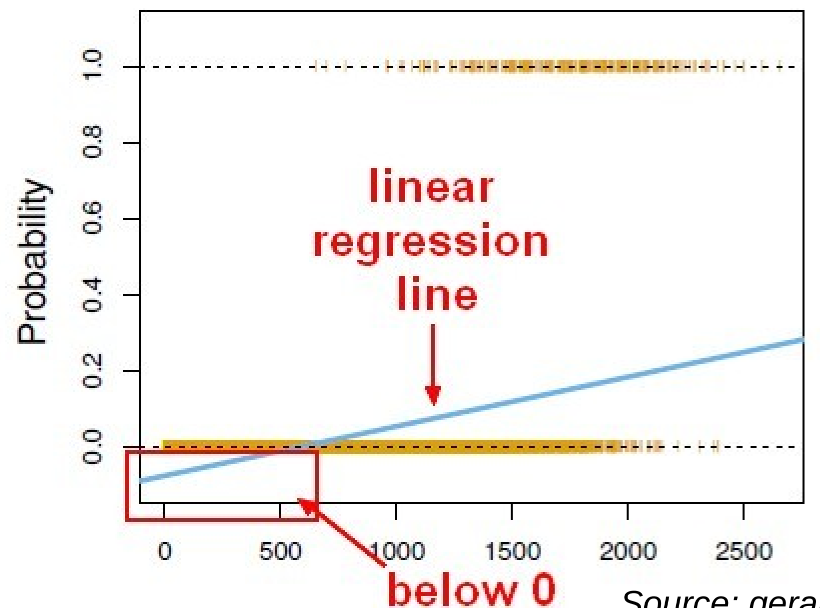
Linear classification: (Multinomial) logistic regression

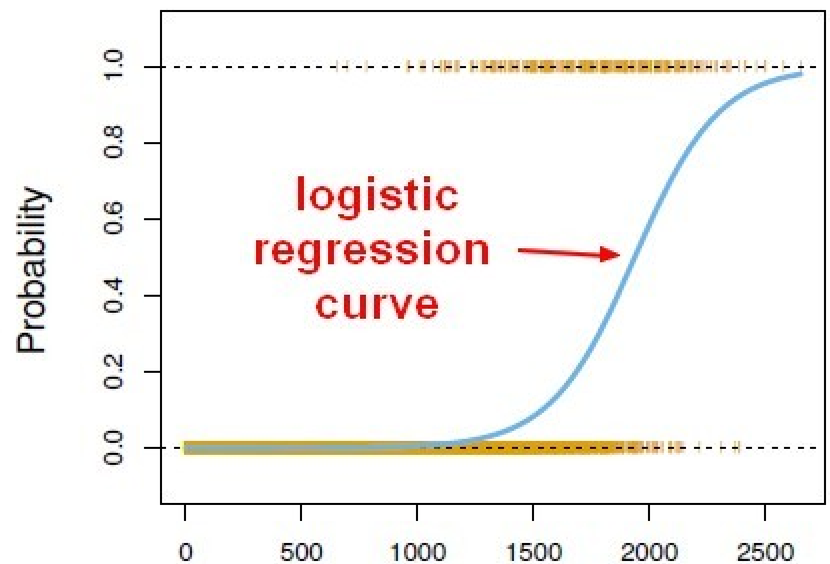
- Assign a weight to each feature for each class: vectors of weights w_c for class c
  - For each class compute $w_c^T x$
  - Add non-linearity $f_c = \exp(w_c^T x)$
  - Normalize it to estimate probabilities: $p(y=c|x) \sim f_c / \sum_{c' \in C} f_{c'}$
- Output value $y = \mathrm{argmax}_c \; p(c|x)$
- Task for linear classification:
  - Given labeled examples
  - Find w_c vectors such that the output value is not wrong for as many training examples as possible

Supervised doc. classification: discrete

(multinomial) logistic regression
= softmax classification
≠ Naive Bayes, Support Vector Machine

Why $\exp(w_c^T x)$?





Source: gerardnico.com/wiki/data_mining/simple_logistic_regression

Supervised doc. classification Softmax classification

Estimating parameters

- Choose parameter which minimize error over training data

Loss function J (aka cost f. or objective f.)

Cross-entropy error on one example (x_i,y_i)

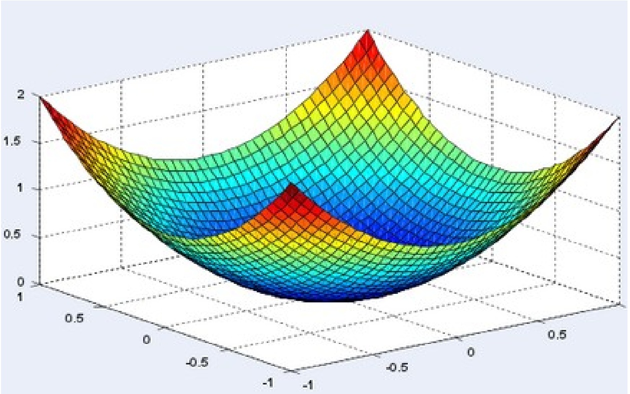
– We want to maximize probability of correct class i.e. **minimize** negative log probability of the correct class

$$J_i(W) = -\log P(y_i = c \mid x_i) = -\log\left(\frac{\exp(W_c x)}{\sum_{c' \in C} \exp(W_{c'} x)}\right)$$

Supervised doc. classification Softmax classification

Estimating parameters

- Search for parameters which minimize error over training data
- **Stochastic gradient descent**
  - Start with random parameters
  - Select K examples (mini-batch) at random
  - Change parameters a little bit *towards minimum of loss function* for those K
  - Continue until loss function converges (or increases)
- Another alternative: analytically calculate optimal parameters



Source: staesthetic.wordpress.com

Supervised doc. Classification
Softmax classification

Stochastic gradient descent

- Start with random parameters: W
- Each epoch
  - Shuffle training data
  - For each mini-batch (set of K examples)
    - Compute the loss function (forward)
    - Compute the gradient of the loss function (backward)
    - Update parameters: (learning rate η)
  - Measure train and dev. accuracy

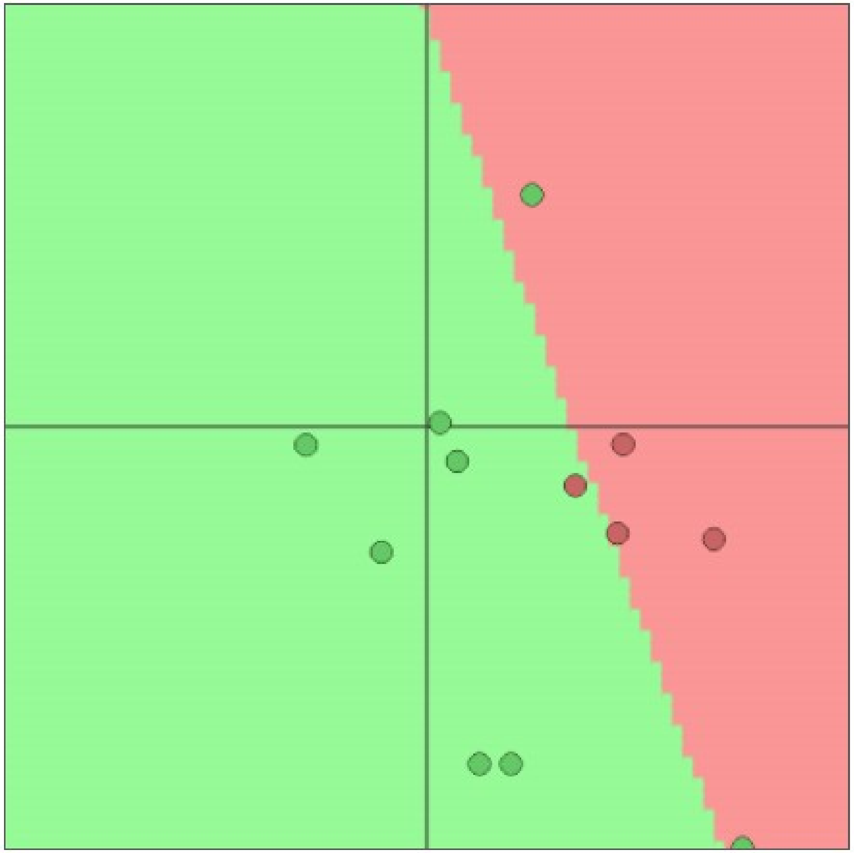
$$W = W - \eta \frac{1}{K} \sum_{i=0}^{K-1} \nabla J_i(W)$$

- Continue until loss function converges / time is up / dev. accuracy stops increasing

Supervised doc. Classification Softmax classification

Intuition:

- Assume 2D vectors
- W defines the linear decision boundary

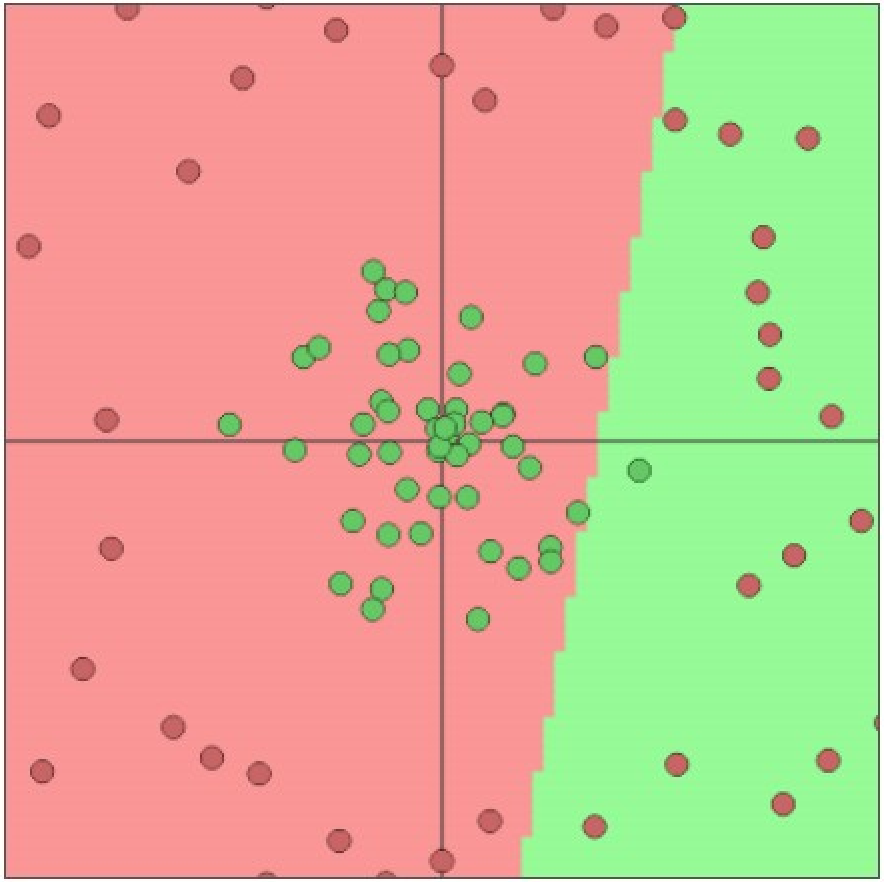


Source: *http://cs.stanford.edu/people/karpathy/convnetjs/demo/classify2d.html*

Supervised doc. Classification
Softmax classification

Intuition:

- Assume 2D vectors
- W defines the linear decision boundary
- Limitations!
  - Kernel trick (SVM)
  - Deep learning



Source: http://cs.stanford.edu/people/karpathy/convnetjs/demo/classify2d.html

Supervised doc. Classification
Softmax classification

Overfitting and regularization

- W can be very good for training, with enough layers and capacity the model can memorize the training data!
  - Generalize very poorly to test data (= the real world)
- First solution: add a regularizer to the loss function that avoids the model to fit the training data

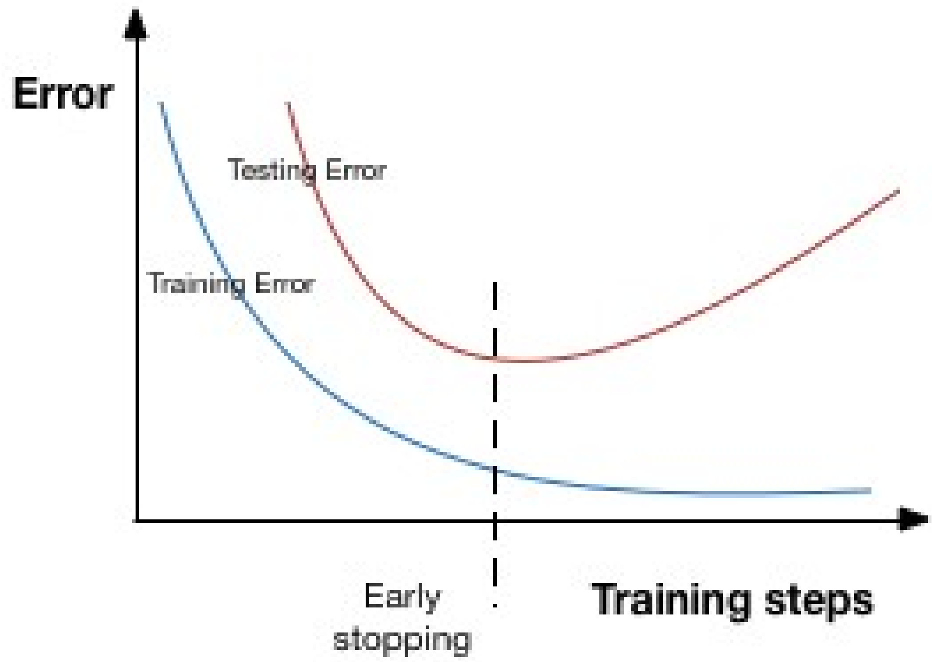
$$J_i(W) = -\log\left(\frac{\exp\left(W_{c_i}^T x\right)}{\sum_{c' \in C} \exp\left(W_{c'}^T x\right)}\right) + \lambda \sum_k W_k^2$$

squared L2 norm

Supervised doc. Classification Softmax classification

Overfitting and regularization

- Overfitting can be seen in this graph
- **Early stopping** finishes training as soon as development error starts to increase



Source: chatbotslife.com

- **Experimental setup**:
  %80 train, %10 development, %10 test (blind!!)
- **Model selection**:
  best accuracy (lowest error) at development

Quiz

Find definition and slide for the following:

- Supervised machine learning
- Document classification
- Document regression
- Linear regression
- Logistic regression
- Softmax classification
- Loss function
- Gradients ∇

Introduction – plan for this session

- Machine learning, Deep learning
- Natural Language Processing
- A sample NLP task with ML
  - sentiment analysis
  - features (bag of words)
  - classification with logistic regression
- **Tensorflow (tomorrow)**

Plan for the course

- 7 theoretical sessions (Eneko, Gorka) (approx. 150 min, time for questions)
- 7 hands-on laboratories (150 minutes)
  - Master students => Onsite (Oier)
  - Independent course => Online (Ander) (the same webex link)
- Calendar with slides and labs: http://ixa2.si.ehu.eus/eneko/dl4nlp
- Note on downloading/uploading the labs data folder => do it before lab, at home

THANKS!

Acknowledgements:

- Overall slides: Sam Bowman (NYU), Chris Manning and Richard Socher (Stanford)
- All source url's listed in the slides