

# CORSO DI *WEB MINING E RETRIEVAL* *- INTRODUZIONE AL WM -*

---

Corso di Laurea in Informatica, Ing. Internet,  
Ing. Informatica, Ing. Gestionale  
(a.a. 2020-2021)

Roberto Basili

# Overview

- Web Mining & Retrieval: Motivazioni e prospettive
  - Web, User-generated contents, Social Media
  - The role of *learning*
  - What is Machine Learning?
  - Data-driven algorithms: sources of complexity
- Main Applications
  - Intelligent Web Search
  - User Profiling for Marketing or Brand reputation management
  - Web Recommending
  - Spoken Dialogue Interaction in Robotics or in Web/mobile Interfaces

# Do you know

There are **100 billion** searches on **Google** every month.



Do you know

If **Facebook** were a country  
It'd be the world's  
**Largest**



Do you know

More than  
**4,000 new books**  
are published every day



Do you know

**Contains more  
information than a  
person was likely to  
come across  
in a lifetime in the  
18th century...**



# Digital snapshot (2019)

Digital 2019 Global Digital Overview (January 2019) v01

**JAN  
2019**

## ANNUAL DIGITAL GROWTH

THE YEAR-ON-YEAR CHANGE IN KEY STATISTICAL INDICATORS

TOTAL  
POPULATION



**+1.1%**

JAN 2018 – JAN 2019

**+84 MILLION**

UNIQUE  
MOBILE USERS



**+2.0%**

JAN 2018 – JAN 2019

**+100 MILLION**

INTERNET  
USERS



**+9.1%**

JAN 2018 – JAN 2019

**+367 MILLION**

ACTIVE SOCIAL  
MEDIA USERS

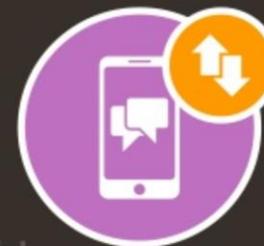


**+9.0%**

JAN 2018 – JAN 2019

**+288 MILLION**

MOBILE SOCIAL  
MEDIA USERS



**+10%**

JAN 2018 – JAN 2019

**+297 MILLION**

**SOURCES:** POPULATION: UNITED NATIONS; U.S. CENSUS BUREAU. MOBILE: GSMA INTELLIGENCE. INTERNET: INTERNETWORLDSTATS; ITU; WORLD BANK; CIA WORLD FACTBOOK; EUROSTAT, LOCAL GOVERNMENT BODIES AND REGULATORY AUTHORITIES; MIDEASTMEDIA.ORG; REPORTS IN REPUTABLE MEDIA. SOCIAL MEDIA: PLATFORMS' SELF-SERVE ADVERTISING TOOLS; PRESS

 **Hootsuite**  **we are social**

# Digital snapshot Europe (2019)

Digital 2019 Global Digital Overview (January 2019) v01

**JAN  
2019**

## DIGITAL IN EUROPE IN 2019

THE ESSENTIAL HEADLINE DATA YOU NEED TO UNDERSTAND MOBILE, INTERNET, AND SOCIAL MEDIA USE

TOTAL  
POPULATION



**846.0**  
MILLION

URBANISATION:

**74%**

MOBILE  
SUBSCRIPTIONS



**1.101**  
BILLION

vs. POPULATION:

**130%**

INTERNET  
USERS



**724.7**  
MILLION

PENETRATION:

**86%**

ACTIVE SOCIAL  
MEDIA USERS

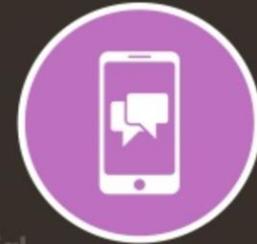


**462.5**  
MILLION

PENETRATION:

**55%**

ACTIVE MOBILE  
SOCIAL USERS



**393.4**  
MILLION

PENETRATION:

**47%**

18

**SOURCES:** POPULATION: UNITED NATIONS; U.S. CENSUS BUREAU. MOBILE: GSMA INTELLIGENCE. INTERNET: INTERNETWORLDSTATS; ITU; WORLD BANK; CIA WORLD FACTBOOK; EUROSTAT, LOCAL GOVERNMENT BODIES AND REGULATORY AUTHORITIES; MIDEASTMEDIA.ORG; REPORTS IN REPUTABLE MEDIA. SOCIAL MEDIA: PLATFORMS' SELF-SERVE ADVERTISING TOOLS; PRESS RELEASES AND INVESTOR EARNINGS ANNOUNCEMENTS; ARAB SOCIAL MEDIA REPORT; TECHRASA; NIKU AGHAIE; ROSEKU. \*ALL LATEST AVAILABLE DATA IN JANUARY 2019.

18 of 221



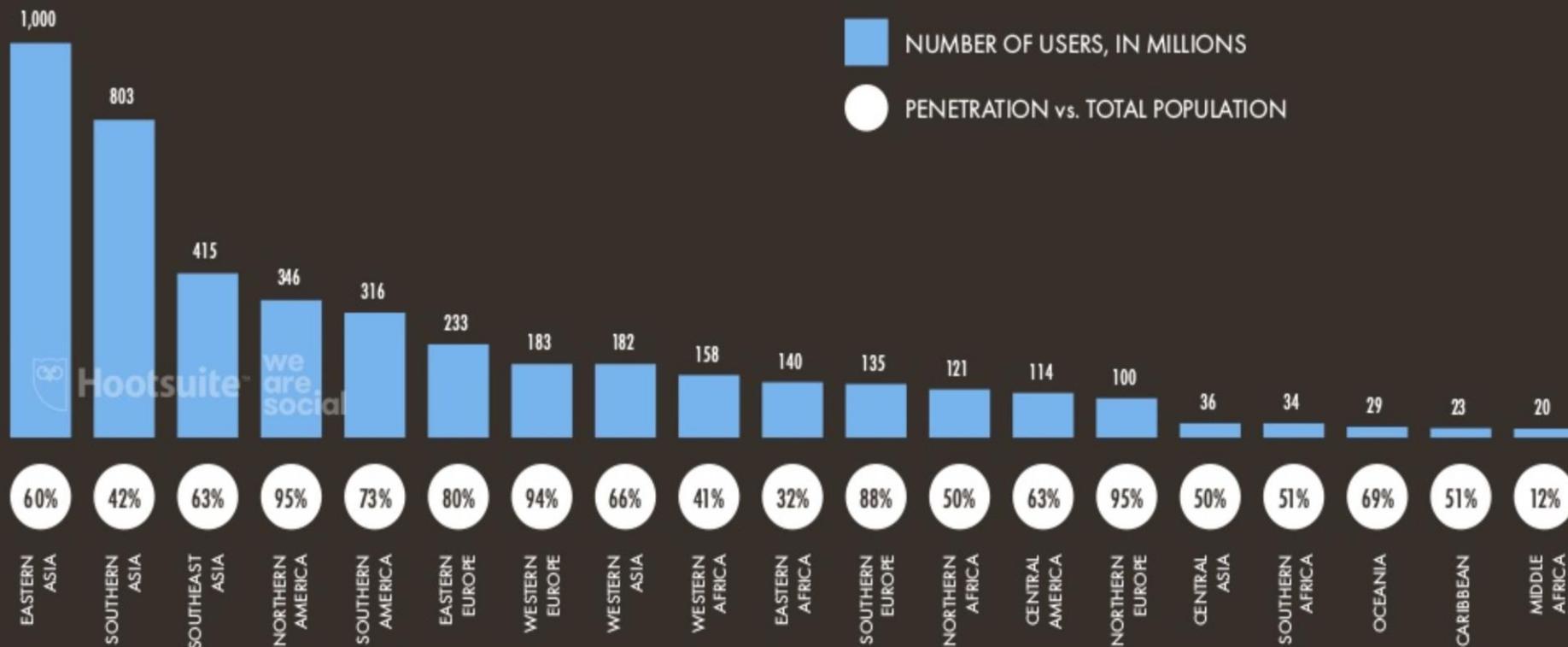
Hootsuite™

we  
are  
social  
in SlideShare

JAN  
2019

# INTERNET USE: REGIONAL OVERVIEW

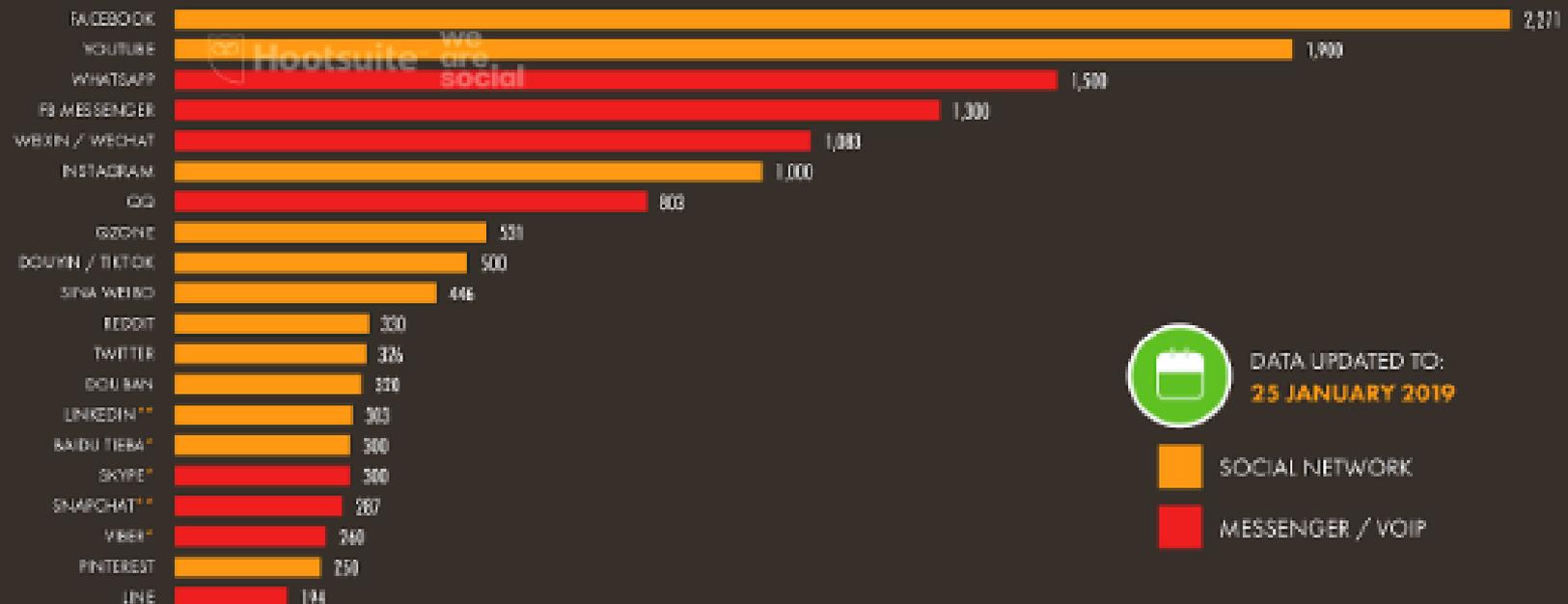
INTERNET USERS IN MILLIONS, AND INTERNET PENETRATION vs. TOTAL POPULATION, BY GLOBAL REGION



JAN  
2019

## SOCIAL PLATFORMS: ACTIVE USER ACCOUNTS

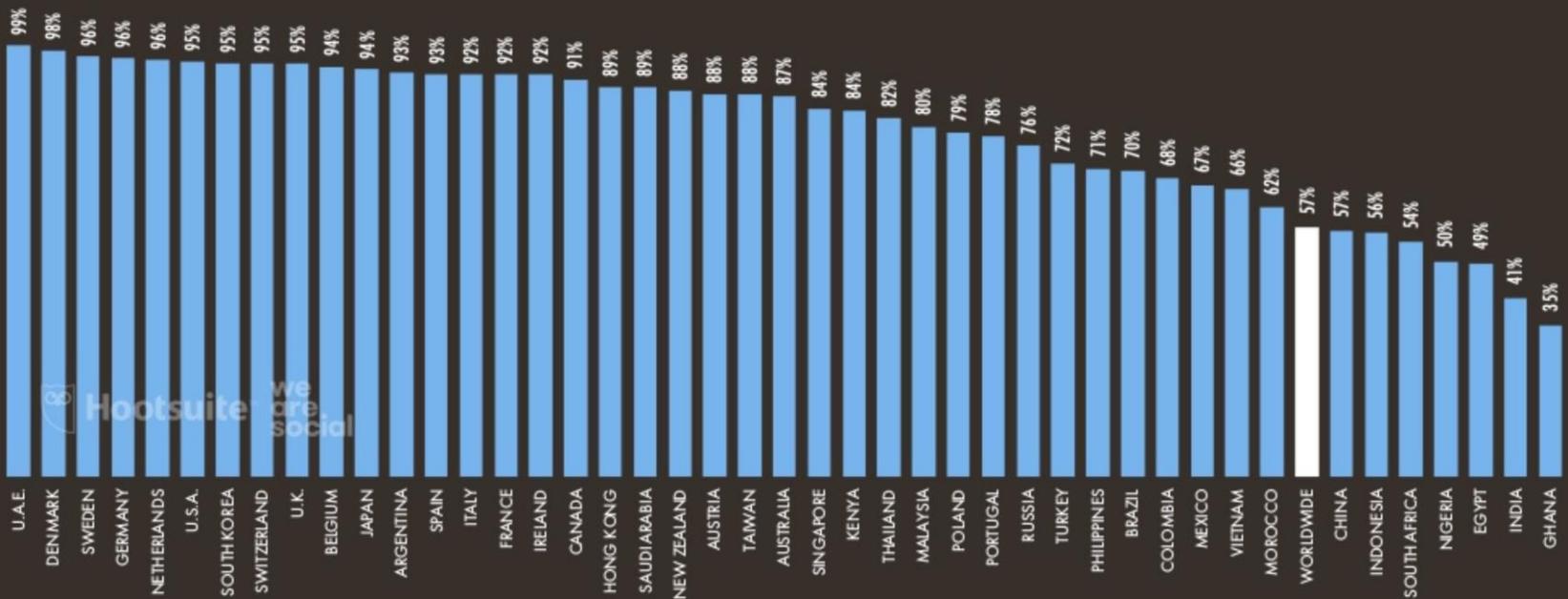
BASED ON MONTHLY ACTIVE USERS, USER ACCOUNTS, OR UNIQUE VISITORS TO EACH PLATFORM, IN MILLIONS



JAN  
2019

## INTERNET PENETRATION IN 2019

PERCENTAGE OF THE TOTAL POPULATION THAT USES THE INTERNET



35

SOURCES: INTERNETWORLDSTATS, ITU; WORLD BANK; CIA WORLD FACTBOOK; LOCAL GOVERNMENT BODIES AND REGULATORY AUTHORITIES; MIDEASTMEDIA.ORG; REPORTS IN REPUTABLE MEDIA. NOTE: PENETRATION FIGURES ARE BASED ON TOTAL POPULATION, REGARDLESS OF AGE.

JAN  
2019

## WORLD'S MOST VISITED WEBSITES (SIMILARWEB)

SIMILARWEB'S RANKING OF THE WORLD'S MOST VISITED WEBSITES, BASED ON TOTAL GLOBAL WEBSITE TRAFFIC

#	WEBSITE	CATEGORY	TIME PER VISIT
01	GOOGLE.COM	SEARCH	09M 12S
02	YOUTUBE.COM	VIDEO	21M 36S
03	FACEBOOK.COM	SOCIAL	11M 44S
04	BAIDU.COM	SEARCH	06M 53S
05	WIKIPEDIA.ORG	REFERENCE	03M 45S
06	YAHOO.COM	PORTAL	06M 26S
07	TWITTER.COM	SOCIAL	09M 14S
08	PORNHUB.COM	ADULT	10M 16S
09	YANDEX.RU	SEARCH	10M 43S
10	INSTAGRAM.COM	SOCIAL	06M 25S

#	WEBSITE	CATEGORY	TIME PER VISIT
11	AMAZON.COM	SHOPPING	06M 18S
12	XVIDEOS.COM	ADULT	12M 34S
13	XNXX.COM	ADULT	14M 39S
14	AMPPROJECT.ORG	NEWS	03M 53S
15	LIVE.COM	EMAIL	07M 15S
16	VK.COM	SOCIAL	16M 50S
17	NETFLIX.COM	VIDEO	09M 14S
18	QQ.COM	PORTAL	04M 00S
19	MAIL.RU	PORTAL	07M 38S
20	REDDIT.COM	SOCIAL	09M 13S

JAN  
2019

## E-COMMERCE DETAIL: CONSUMER GOODS

OVERVIEW OF THE E-COMMERCE MARKET FOR CONSUMER GOODS, WITH VALUES IN U.S. DOLLARS

TOTAL NUMBER OF PEOPLE  
PURCHASING CONSUMER  
GOODS VIA E-COMMERCE



**2.818**  
BILLION

YEAR-ON-YEAR CHANGE

**+3.1%**

PENETRATION OF CONSUMER  
GOODS E-COMMERCE  
(TOTAL POPULATION)



**37%**

statista

VALUE OF THE CONSUMER  
GOODS E-COMMERCE MARKET  
(TOTAL ANNUAL SALES REVENUE)



**\$1.786**  
TRILLION

YEAR-ON-YEAR CHANGE:

**+14%**

we  
are  
social

AVERAGE ANNUAL REVENUE  
PER USER OF CONSUMER  
GOODS E-COMMERCE (ARPU)



**\$634**

YEAR-ON-YEAR CHANGE:

**+11%**

JAN  
2019

## LITERACY RATE BY GENDER AND REGION

PERCENTAGE OF EACH REGION'S POPULATION AGED 15 AND ABOVE WHO CAN READ AND WRITE, SPLIT BY GENDER



# Dealing with *real* Social media data



# WM&R: Motivazioni

- *Cos'è il Web Mining?*
- *Perché IR?*
- *Perché Apprendimento Automatico?*
- *Quale contributo l'IR fornisce alle tecnologie di sfruttamento delle informazioni del Web?*
- *Quali sono le prospettive per l'impiego di tali tecnologie?*

# Cos'è il Web Mining?

- Web Mining attualmente si riferisce ad un insieme di tecnologie necessarie per lo sfruttamento delle *informazioni pubblicamente disponibili nel Web*
  - Contenuti: dati ma anche ... persone, luoghi, eventi, concetti, ...
  - Relazioni:
    - Link strutturali
    - Collegamenti tematici, concettuali e interpersonali
    - Ridondanze/ analogie
  - Risorse On-Line (Wikipedia)
  - Multilingualità
  - Trend e comportamenti collettivi
  - Opinioni

# Perché Information Retrieval?

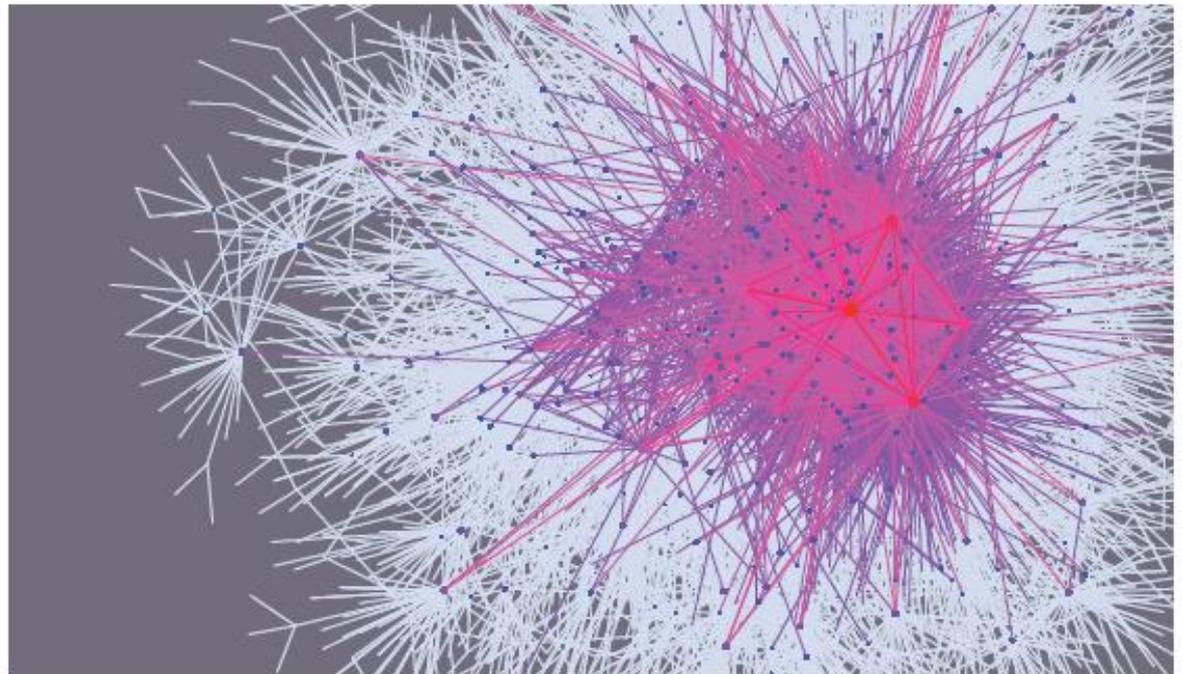
- La taglia delle informazioni in gioco pone il problema della *localizzazione*
- Accedere in modo automatico è possibile solo governando il problema di sapere **dove** si trova una informazione *rilevante*
- La ricerca corrisponde al calcolo di una funzione *aleatoria* di mapping tra requisiti e informazione utile

# Machine Learning vs IR?

- La eterogeneità delle informazioni produce significativi effetti di incertezza nel processo di ricerca riguardo ad aspetti diversi del processo di IR
  - Incompletezza della informazione:
    - Query brevi come informazione (incomplete) sui fabbisogni informativi
  - Ricchezza di dati, formati e modalità di accesso
    - I contenuti sono sparsi in diverse forme nei dati
  - Requisiti vaghi
    - Spesso molte informazioni sono esplicite solo nel contesto
  - Aspetti soggettivi
    - La rilevanza dipende dallo user e non solo dal contenuto
  - Tempestività ed autorevolezza

# Machine Learning vs. IR

- La pervasività degli elementi di incertezza rende impraticabile la ricerca di soluzioni esaustive (ottimi globali)
- “*Finding diamonds in the rough*”  
(Fan Chung, UCSD)



# Machine Learning vs. IR

- Le tecniche di ML propongono una ampia serie di algoritmi, strategie e tecniche per la produzione di soluzioni *sub-ottime* ma efficaci
- Nel processo di *learning* i dati suggeriscono la ipotesi risolutiva per la funzione di *mapping*
- Tale ipotesi è attesa migliorare la prestazione complessiva del sistema di base
  - Accuratezza
  - Efficienza computazionale

# Machine Learning

- (Langley, 2000): l'Apprendimento Automatico si occupa dei meccanismi attraverso i quali un agente intelligente migliora nel tempo le sue prestazioni  $P$  nell'effettuare un compito  $C$ .
- La prova del successo dell'apprendimento è quindi nella capacità di misurare l'incremento  $\Delta P$  delle prestazioni sulla base delle esperienze  $E$  che l'agente è in grado di raccogliere durante il suo ciclo di vita.
- La natura dell'apprendimento è quindi tutta nella caratterizzazione delle nozioni qui primitive di *compito*, *prestazione* ed *esperienza*.

# Esperienza ed Apprendimento

- L'esperienza, per esempio, nel gioco degli scacchi può essere interpretata in diversi modi:
  - i dati sulle vittorie (e sconfitte) pregresse per valutare la bontà (o la inadeguatezza) di strategie e mosse eseguite rispetto all'avversario.
  - valutazione fornita sulle mosse da un docente esterno (oracolo, guida).
  - Adeguatezza dei comportamenti derivata dalla auto-osservazione, cioè dalla capacità di analizzare partite dell'agente contro se stesso secondo un modello esplicito del processo (partita) e della sua evoluzione (comportamento, vantaggi, ...).

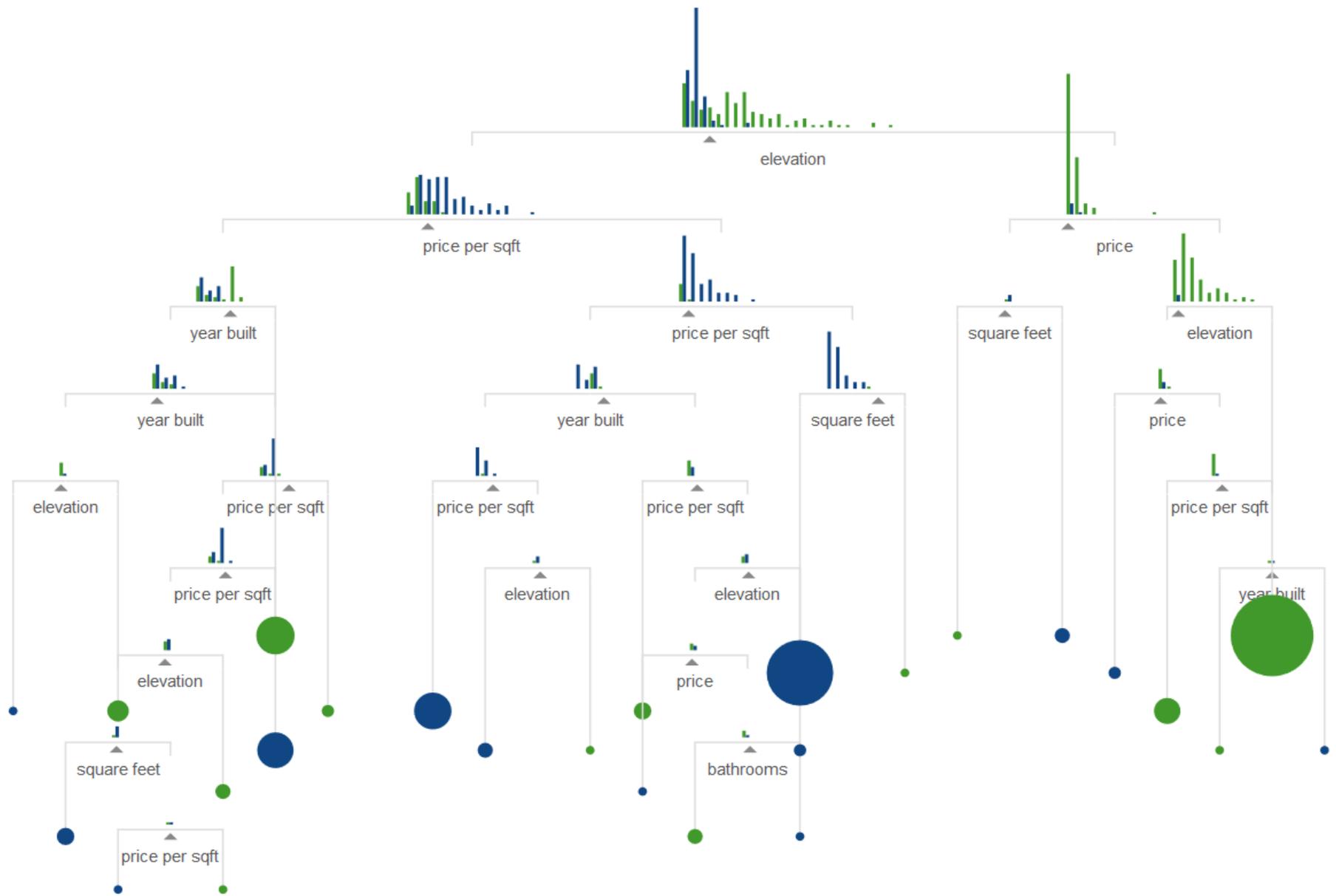
# ML: una introduzione visuale

- See URL: [http://www.r2d3.us/visual-intro-to-machine-learning-part-1/?imm\\_mid=0d76b4&cmp=em-data-na-na-newsltr\\_20150826](http://www.r2d3.us/visual-intro-to-machine-learning-part-1/?imm_mid=0d76b4&cmp=em-data-na-na-newsltr_20150826)

# Apprendimento e Classi di Algoritmi

- Acquisizione di:
  - Funzioni logiche booleane, (ad es., alberi di decisione)
  - Induzione: determinazione ricorsiva delle CNES che caratterizzano i diversi sottogruppi .
- Approcci probabilistici:
  - Funzione target di Probabilità, (ad es., classificatore Bayesiano)
  - Induzione: Stima delle probabilità (in quanto parametri).
- Approcci geometrici
  - Funzioni di separazione in spazi vettoriali (lineari e non)
    - KNN
    - Funzioni Lineari, perceptroni, Neural Networks, Support Vector Machines,...
    - Embeddings, analisi spettrale (trasformazioni di spazio)
  - Induzione: parametrizzare la funzione appartenente ad una certa classe (ad es. polinomi di grado  $n$ )

# Es. apprendimento alberi di decisione

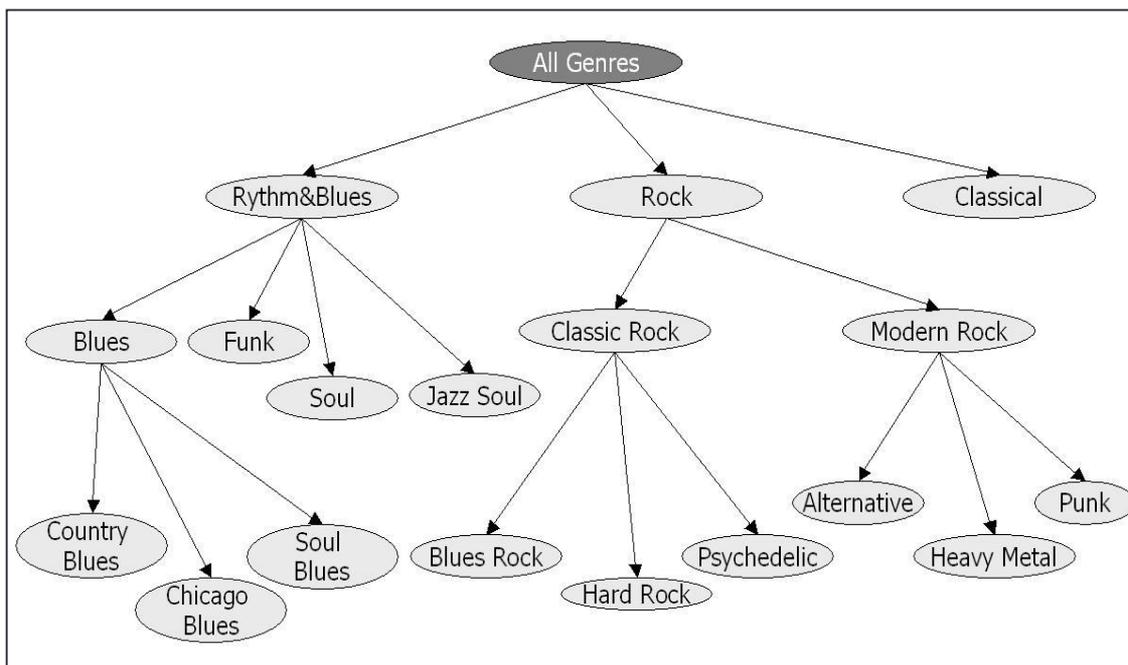


# Apprendimento senza supervisione

- In assenza di un oracolo o di conoscenze sul task esistono ancora molti modi di migliorare le proprie prestazioni, ad es.
  - Migliorando il proprio modello del mondo (acquisizione/*discovery* della conoscenza)
  - Migliorando le proprie prestazioni computazionali (ottimizzazione)

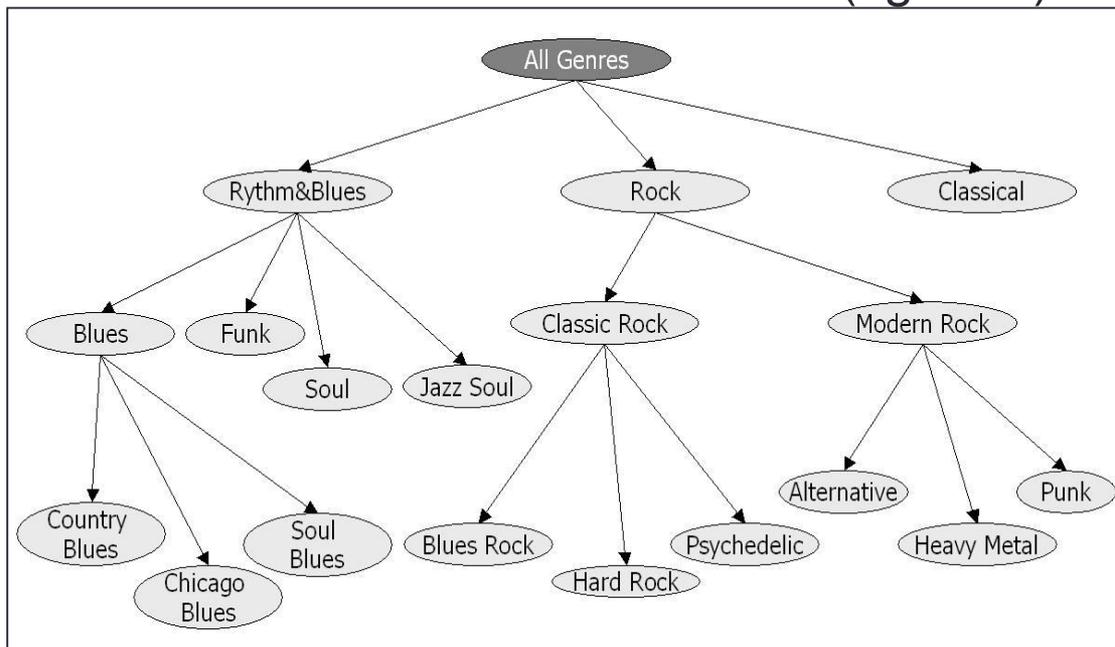
# Apprendimento senza supervisione

- Esempio:
  - una collezione *mp3* può essere organizzata in generi attraverso il raggruppamento di brani simili secondo proprietà audio (*clustering*): tale organizzazione è naturalmente gerarchica
  - Il miglioramento avviene quindi almeno rispetto agli algoritmi di ricerca: la organizzazione gerarchica consente di esaminare solo i membri dell'insieme in alcune classi (i generi).



# Apprendimento senza supervisione

- Esempio: Al termine del processo di acquisizione il sistema dispone di un sistema di classi e relazioni indotti che migliora la sua interazione futura con l'ambiente operativo (ad es. l'utente)
- Il miglioramento avviene quindi almeno rispetto agli algoritmi di ricerca: la organizzazione gerarchica consente di esaminare solo i membri dell'insieme in alcune classi (i generi).





# Prospettive delle tecnologie WM&R

- Crescita esponenziale della taglia dei problemi
- Crescente interesse verso processi di IR agenti su dati complessi (multimediali, sociali)
- Web partecipativo: dal Web 2.0 in poi
- Ruolo crescente della mediazione degli strumenti informatici
  - Software as a Service
  - Personalizzazione
  - Digital Transformation, Industry 4.0
- Big Data challenges:
  - Scala dei problemi e scala del processo di elaborazione
  - Opacità Semantica: *data* *vz.* *interpretazione*
  - Interoperabilità

# Information, Web and the language

Web contents, characterized by rich multimedia information, are mostly **opaque from a semantic standpoint**

Today is 2011年11月13日 星期日 顯示器最佳分辨率 1024X768 今日天氣 加入最愛 設為首頁 大公網新版

www.takungpao.com.hk

2011 中国証券金紫荆獎 Golden Bauhinia Awards

首頁 國內 國際 港澳 兩岸 評論 財經 體育 教育 科技 醫學 娛樂 文化 副刊 軍事 生活 旅遊 圖片 博客

關鍵詞: 欄目: 全部 最近三個月 三個月之前 檢索

▶ 手機新聞 ▶ 手機博客 ▶ 漢語學習 ▶ 新聞點擊排行

滾動新聞:

**胡總語特首:防範經濟金融風險**  
胡錦濤在夏威夷會見出席APEC峰會的曾蔭權。他祝賀香港區議會選舉成功,並充分肯定曾蔭權及港府工作,要求做好經濟金融風險防範

**胡連會登場 共同宣示九二共識**  
胡錦濤第四次在APEC峰會期間會見連戰。他強調,認同「九二共識」是兩岸開展對話協商的必要前提,也是兩岸關係和平發展的重要基礎

西藏黨代會高調反「藏獨」 德國作家:外埠雜誌報道西藏  
傳媒入日本福島核電站探險 英國大裁軍 傷兵難雜免  
滇礦難已30死 13人生還 礦工講述內幕 事故並不意外  
范徐麗泰認民望跌最不耐 選委再獲60提名表 累積逾千人  
聖保羅中學本月底截止招新 選委再獲60提名表 累積逾千人  
民調逆轉 藍高層:國親吵鬧地 秋門訴求多 向藍綠表不滿  
世界新七奇觀 亞洲景佔四席 新奇觀選舉惹爭議

中國實體書店苦苦掙扎求存 加入TPP 台密集會談探險  
香港人家/蔡仕榮 人生導師 活出自我 我香港人家/教導子女...  
債務危機好 港ADR幾全線造藍 歐元反降 兌美元逼近1.38  
入世十年/充分對接 華強北最難 入世十年/挑戰「二次」...  
抽身除「雜軍」 工人險生國難 南亞漢命案 警日籍妻

2011APEC 港黑金事件 2011  
神八天宮對接 第七次陳江會 李...  
9.1衝擊事件 中國航母試航 辛...

http://www.takungpao.com.hk/news/11/11/13/2011\_apec\_xgbd-1423309.htm

# Information, Web and language

Hu meets KMT honorary chairman in Hawaii - People's Daily Online - Mozilla Firefox

File Modifica Visualizza Cronologia Segnalibri Yahoo! Strumenti Aiuto

Hu meets KMT honorary chairman

Indietro Avanti Download

*Chinese President Hu Jintao (R) shakes hands with Honorary Chairman of the Chinese Kuomintang (KMT) Lien Chan, in Honolulu, Hawaii, the U.S., Nov. 11, 2011.*  
*(Xinhua/Huang Jingwen)*

HONOLULU, United States, Nov. 11 (Xinhua) -- Hu Jintao, general secretary of the Central

Latest News: • Indonesia to host European Higher Education Fair

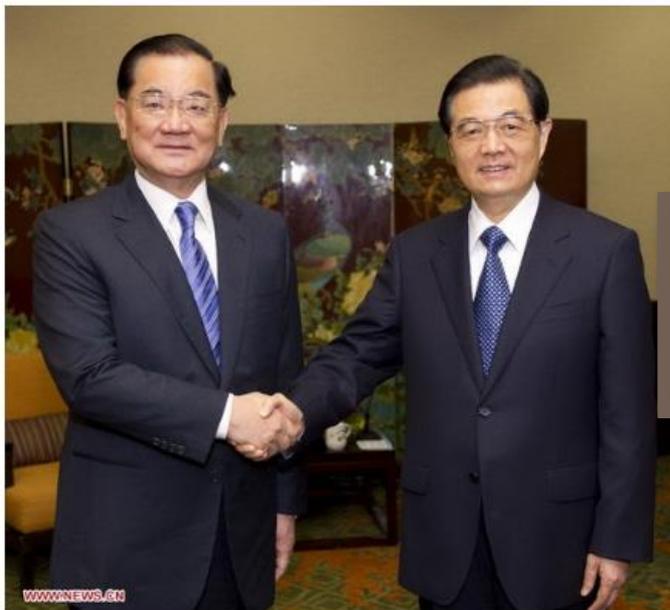
Beijing Sunny 15 / 1 City Forecast

Home >> China Politics

## Hu meets KMT honorary chairman in Hawaii

(Xinhua)

11:10, November 12, 2011 🔍 +-



*Chinese President Hu Jintao (R) shakes hands with Honorary Chairman of the Chinese Kuomintang (KMT) Lien Chan, in Honolulu, Hawaii, the U.S., Nov. 11, 2011.*

Selections for you



Miao ethnic group celebrates Miao's New Year in SW China



World's first Angry Birds exclusive shop opens in Helsinki

*Who is Hu Jintao?*

Most Popular

- 1 Hu reaffirms support to Hong Kong's sta...
- 2 Hu meets KMT honorary chairman in Hawaii
- 3 China in APEC: a mutually beneficial en...
- 4 Night life in Shanghai
- 5 China's 2011 foreign trade to grow 20 p...
- 6 Beijing house prices stumble 5.1 pct as...
- 7 Lama students start school in Tibet Col...
- 8 Police in central China crack phoney ca...



Hu Jintao



Ricerca

Circa 725.000 risultati (0,09 secondi)

- Tutto
- Immagini
- Mappe
- Video
- Notizie
- Shopping
- PIÙ conte

Tutti i ri  
Per argomento

Qualsiasi dimensione

- Grandi
- Medie
- Icone
- Maggiori di...
- Dimensioni esatte...

Qualsiasi colore

- A colori
- Bianco e nero



Qualsiasi tipo

- Volti
- Foto
- Clip art
- Disegni

Visual standard

Mostra dimensioni

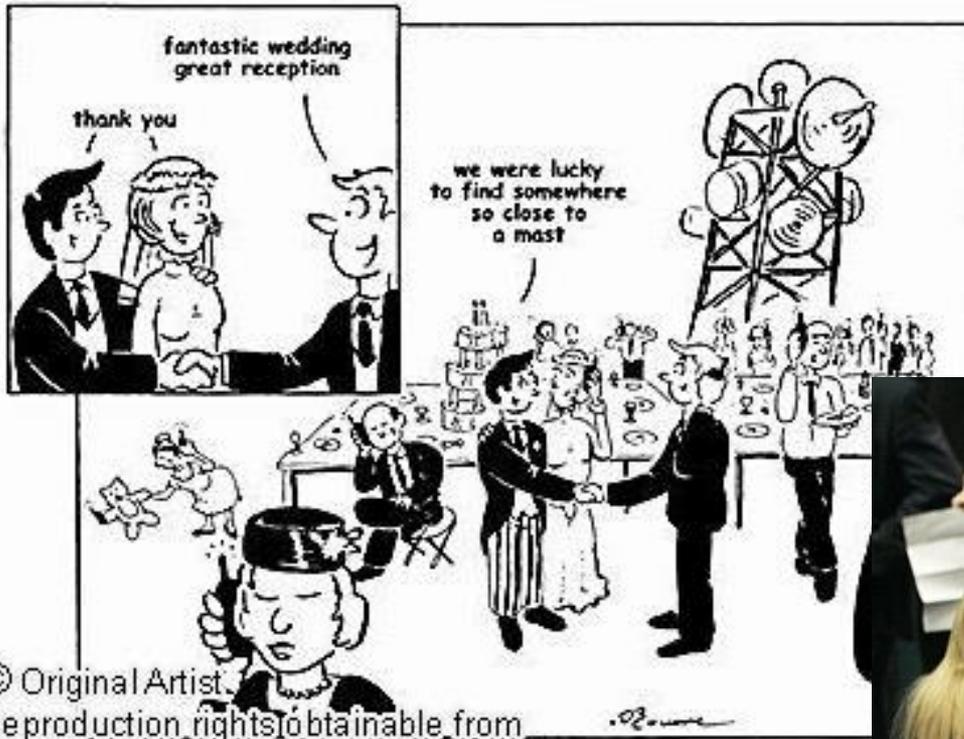


# Content Semantics and Natural Language

- Human languages are the main carrier of the information involved in processes such as *retrieval*, *publication* and *exchange* of knowledge as it is associated to the open Web contents
- Words and NL syntactic structures express concepts, activities, events, abstractions and conceptual relations we usually share through data
- “*Language is parasitic to knowledge representation languages but the viceversa is not true*” (Wilks, 2001)

# NL and Knowledge

Natural Languages are even too successful in modern communication



search ID



© Original Artist.  
Reproduction rights obtainable from  
[www.CartoonStock.com](http://www.CartoonStock.com)

© AFP/Getty Images

# Natural Languages: undesirable aspects



Some crucial language understanding problems affect the comprehension of human speakers:

- **Noise**
- **Incompleteness**
- **Ambiguity**

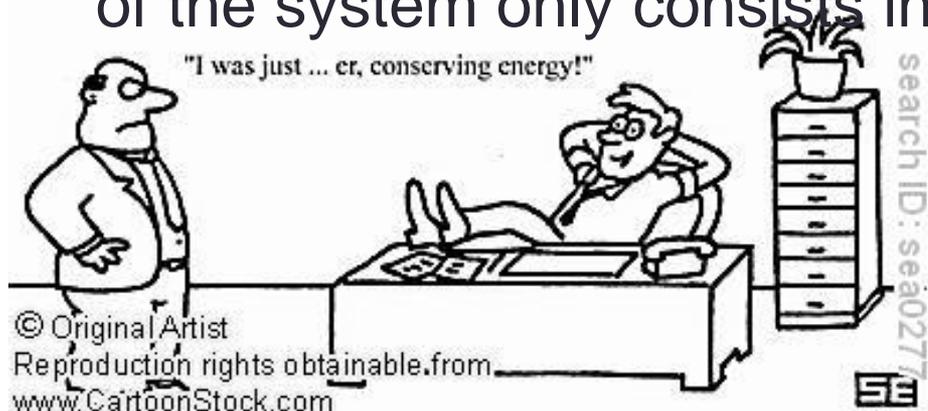
... and for machines and software systems they are even more critical



- Since '90s successful methods inspired by data-driven paradigms allow a quantitative (i.e. numerical) approach:
  - Markov models of languages and communication processes
  - Neural nets approaches (annotated examples to induce and optimize robust and large scale models)

# Benefits of a data-driven approach

- Very effective learning algorithms available (e.g. Support Vector Machines, Neural Networks)
- The ML technology is *portable* while imperative coding is *task (i.e. scenario) specific*
- Very accurate solutions can be obtained
- Gathering training data *much less expensive* than rule coding
- In dynamically evolving scenarios, incremental refinement of the system only consists in re-training



# Data Mining: perspectives and benefits

- Technical advantages
  - Self adaptivity to changing operational conditions (i.e. domain)
  - Better SW management and incremental maintenance
  - More flexibility for special-purpose versioning:
    - No need for re-engineering or independent software developments
    - Just new domain-specific examples are needed
- Cost benefits
  - The data-driven approach has been shown to reduce the development costs up to 80-90% in several NLP tasks
- Market benefits
  - Reduced time-to-market
  - Competitive advantages: the lack of similar products makes the system targeted strongly competitive solutions

# Semantics and News

Applicazioni Risorse Sistema mar 27 lug, 23.47 dan

Gmail ... x SRL\_EN x Come ... x R Econo... x Googl... x Tanl It... x Frame... x SRL\_EN x Econo... x

file:///home/danilo/Downloads/SRL\_ITA/sorgente/Economia%20-%20Repubblica.it.html

Telefilm in stream... Telefilm in stream... Flash Forward pri... Telefilm in stream... Cronologia Altri Pr



L'ad punta a nuove regole sulla base del modello Pomigliano. L'annuncio, che prevede l'uscita da Federmeccanica, domani al vertice con il governo o giovedì con una lettera a Bombassei. Potrebbe avvenire assieme alla decisione di creare una new company per

Pomigliano di SALVATORE TROPEA

**Cisl-Uil: "L'accordo di categoria non si tocca"** di S. PAROLA

**Sacconi: "Su Fiat partita aperta"**

**Nasce Fabbrica Italia Pomigliano**

## Si dimette il capo di Bp buonuscita un milione di sterline



Oggi l'annuncio: a Tony Hayward subentrerà il direttore esecutivo Robert Dudley. **I costi legati al disastro sono saliti a 32,2 miliardi di dollari**, ma la società li detraerà evitando di versare al fisco Usa 10 miliardi

## Manager Usa, è Ellison di Oracle il più pagato del decennio



Ha guadagnato 1,84 miliardi di dollari. Nella classifica del *Wall Street Journal* sui leader delle società quotate, secondo con 1,14 miliardi il capo di Expedia, terzo Irani di Occidental Petroleum. Solo quarto Steve Jobs

**Il nemico alle porte**

**La Consob e la mano invisibile**

Altri articoli

**PICCOLE GRANDI IMPRESE**  
DI LUCA PAGNI

**La grande sfida del teleshopping**

**La crisi colpisce anche i porti turistici ma siamo sicuri che sia un male?**

Altri articoli

**PERCENTUALMENTE**  
DI ROSARIA AMATO

**La prova del 9**

**L'export risolve il Pil, ma non le famiglie**

Altri articoli

### GLI ESPERTI RISPONDONO

**CASA**  
A cura di Antonella Donati

**Compenso extra, quando ne ha diritto l'amministratore**

Mia moglie ed il fratello sono proprietari di un appartamento in condominio. Allo stato

Il tuo libro arriva dove  
hai sempre sognato.

ilmiolibro.it

**24ORE AGI**

**Roma 19:04**  
ACEA: NEL I SEMESTRE UTILE NETTO +52,1% A 2010, MLN

**Parigi 18:42**  
AIR FRANCE-KLM: TORNA IN UTILE NEL PRIMO TRIMESTRE

← 3 → Le altre not

**CREDITO ALLE IMPRESE**

**Microimprese: con la crisi aumenta il rischio di credito**

IN COLLABORAZIONE CON

# Laboratori del Corso

- Nell'ambito dei Laboratori agli studenti saranno resi disponibili:
  - Piattaforme di Machine Learning: Weka, , KERP, TensorFlow, SciKit
  - Strumenti di AI per l'elaborazione dei testi:
    - Recursive Neural Networks per l'apprendimento di lessici vettoriali
    - Parser grammaticali di linguaggi naturali (ita,eng)
    - Ambienti di Enterprise Semantic Search su Web

# Kelp: Java-based kernel framework

The screenshot shows the website for the Semantic Analytics Group @ Uniroma2. The page features a navigation menu with links for People, Research, Teaching, Publications, Projects, Demo & Software, and Contacts. Below the navigation is a large image of a glowing blue brain with circuit-like patterns. The main content area is titled "KeLP (Kernel-based Learning Platform)" and contains the following text:

**KeLP** (Kernel-based Learning Platform) is a machine learning platform developed within the SAG group. It is entirely written in Java and it is strongly focused on *Kernel Machines*. It includes different Online and Batch Learning and Classification algorithms, Kernel functions, ranging from vector-based to structural kernels. **KeLP** allows to build complex kernel machine based systems, leveraging on the Java language and on a JSON interface to store and load classifiers configurations as well as to save the models to be reused.

For a deeper look, you can visit [What's inside KeLP page](#).

**Downloads**

**KeLP** is released under [Maven](#). To use it, please refer to the [Installation](#) page

To download **KeLP** source code you can go to the github [KeLP page](#).

On the right side of the page, there are sections for "Authentication" with a "Log In" link, and "News" with a list of recent updates:

- [SAG's KeLP team ranked first at the SemEval 2016 Community Question Answering Task](#) February 16, 2016
- [KeLP 2.0.2 released!](#) February 16, 2016
- [KeLP 2.0.1 released](#) January 13, 2016
- [The ECIR 2016 paper has been accepted!](#) December 30, 2015
- [KeLP 2.0.0 released](#) December 4, 2015
- [SAG with Reveal @ Maker Faire 2015, Rome!!](#) October 16, 2015

`https://github.com/SAG-KeLP`

`http://sag.art.uniroma2.it/demo-software/kelp/`

# KELP applications: cQA

General Description

Subtasks

Data and Tools

Important Dates

Results

Call for Papers

## SemEval-2016 Task 3

### Task 3: Community Question Answering

Building on the success of [SemEval 2015 Task 3](#) "Answer Selection in Community Question Answering" (see [the task description paper](#)), we propose an extension, which covers a full task on Community Question Answering (CQA) and which is, therefore, closer to a real application (see, e.g., [Qatar Living forum](#)).

CQA systems are gaining popularity online. Such systems are seldom moderated, quite open, and thus they have little restrictions, if any, on who can post and who can answer a question. On the positive side, this means that one can freely ask any question and expect some good, honest answers. On the negative side, it takes effort to go through all possible answers and to make sense of them. For example, it is not unusual for a question to have hundreds of answers, which makes it very time-consuming for the user to inspect and to winnow through them all. The present task could help to automate the process of finding good answers to new questions in a community-created discussion forum (e.g., by retrieving similar questions in the forum and by identifying the posts in the answer threads of those similar questions that answer the original question well).

In essence, the main CQA task can be defined as follows:

*"given (i) a new question and (ii) a large collection of question-comment threads created by a user community, rank the comments that are most useful for answering the new question"*

#### Contact Info

#### Organizers

- Preslav Nakov, Qatar Computing Research Institute, HBKU
- Lluís Màrquez, Qatar Computing Research Institute, HBKU
- Alessandro Moschitti, Qatar Computing Research Institute, HBKU
- Walid Magdy, Qatar Computing Research Institute, HBKU
- James Glass, CSAIL-MIT
- Bilal Randeree, Qatar Living

email : [semeval-cqa@googlegroups.com](mailto:semeval-cqa@googlegroups.com)

#### Other Info

#### Announcements

# KELP applic

General Description

## SemEval-20

Team ID	Team Affiliation
ConvKN	Qatar Computational
ECNU	East China Normal University
ICL00	Institute of Computing Technology, Chinese Academy of Sciences
ICRC-HIT	Intelligence and Information Technology Center, Harbin Institute of Technology
ITNLP-AiKF	Intelligence and Information Technology Center, Harbin Institute of Technology
Kelp	University of Cambridge
MTE-NN	Qatar Computational
overfitting	University of Cambridge
PMI-cool	Sofia University "St. Kliment Ohridski", Sofia
QAIIT	IIT Hyderabad
QU-IR	Qatar University
RDI.team	RDI Egypt, Cairo
SemanticZ	Sofia University "St. Kliment Ohridski", Sofia
SLS	MIT Computational Science and Technology Laboratory
SUper.team	Sofia University "St. Kliment Ohridski", Sofia
UH-PRHLT	Pattern Recognition and Computer Vision Group, Universitat Politècnica de Catalunya
UniMelb	The University of Melbourne
UPC_USMBA	Universitat Politècnica de Catalunya

Tal  
shc  
tea  
the  
rut

	Submission	MAP	AvgRec	MRR	P	R	F1	Acc
<b>1</b>	<b>Kelp-primary</b>	<b>79.19<sub>1</sub></b>	<b>88.82<sub>1</sub></b>	<b>86.42<sub>1</sub></b>	<b>76.96<sub>1</sub></b>	<b>55.30<sub>8</sub></b>	<b>64.36<sub>6</sub></b>	<b>75.11<sub>2</sub></b>
	ConvKN-contrastive1	78.71	88.98	86.15	77.78	53.72	63.55	74.95
	SUper.team-contrastive1	77.68	88.06	84.76	75.59	55.00	63.68	74.50
<b>2</b>	<b>ConvKN-primary</b>	<b>77.66<sub>2</sub></b>	<b>88.05<sub>3</sub></b>	<b>84.93<sub>4</sub></b>	<b>75.56<sub>2</sub></b>	<b>58.84<sub>6</sub></b>	<b>66.16<sub>2</sub></b>	<b>75.54<sub>1</sub></b>
	<b>3 SemanticZ-primary</b>	<b>77.58<sub>3</sub></b>	<b>88.14<sub>2</sub></b>	<b>85.21<sub>2</sub></b>	<b>74.13<sub>4</sub></b>	<b>53.05<sub>10</sub></b>	<b>61.84<sub>8</sub></b>	<b>73.39<sub>5</sub></b>
<b>4</b>	ConvKN-contrastive2	77.29	87.77	85.03	74.74	59.67	66.36	75.41
	<b>4 ECNU-primary</b>	<b>77.28<sub>4</sub></b>	<b>87.52<sub>5</sub></b>	<b>84.09<sub>6</sub></b>	<b>70.46<sub>6</sub></b>	<b>63.36<sub>4</sub></b>	<b>66.72<sub>1</sub></b>	<b>74.31<sub>4</sub></b>
<b>5</b>	SemanticZ-contrastive1	77.16	87.73	84.08	75.29	53.20	62.35	73.88
	<b>SUper.team-primary</b>	<b>77.16<sub>5</sub></b>	<b>87.98<sub>4</sub></b>	<b>84.69<sub>5</sub></b>	<b>74.43<sub>3</sub></b>	<b>56.73<sub>7</sub></b>	<b>64.39<sub>4</sub></b>	<b>74.50<sub>3</sub></b>
	MTE-NN-contrastive2	76.98	86.98	85.50	58.71	70.28	63.97	67.83
	SUper.team-contrastive2	76.97	87.89	84.58	74.31	56.36	64.10	74.34
	MTE-NN-contrastive1	76.86	87.03	84.36	55.84	77.35	64.86	65.93
	SLS-contrastive2	76.71	87.17	84.38	59.45	67.95	63.41	68.13
<b>6</b>	SLS-contrastive1	76.46	87.47	83.27	60.09	69.68	64.53	68.87
	<b>6 MTE-NN-primary</b>	<b>76.44<sub>6</sub></b>	<b>86.74<sub>7</sub></b>	<b>84.97<sub>3</sub></b>	<b>56.28<sub>9</sub></b>	<b>76.22<sub>1</sub></b>	<b>64.75<sub>3</sub></b>	<b>66.27<sub>8</sub></b>
<b>7</b>	<b>SLS-primary</b>	<b>76.33<sub>7</sub></b>	<b>87.30<sub>6</sub></b>	<b>82.99<sub>7</sub></b>	<b>60.36<sub>8</sub></b>	<b>67.72<sub>3</sub></b>	<b>63.83<sub>6</sub></b>	<b>68.81<sub>7</sub></b>
	ECNU-contrastive2	75.71	86.14	82.53	63.60	66.67	65.10	70.95
	SemanticZ-contrastive2	75.41	86.51	82.52	73.19	50.11	59.49	72.26
<b>8</b>	ICRC-HIT-contrastive1	73.34	84.81	79.65	63.43	69.30	66.24	71.28
	<b>8 ITNLP-AiKF-primary</b>	<b>71.52<sub>8</sub></b>	<b>82.67<sub>9</sub></b>	<b>80.26<sub>8</sub></b>	<b>73.18<sub>5</sub></b>	<b>19.71<sub>12</sub></b>	<b>31.06<sub>12</sub></b>	<b>64.43<sub>9</sub></b>
	ECNU-contrastive1	71.34	83.39	78.62	66.95	41.31	51.09	67.86
<b>9</b>	<b>9 ICRC-HIT-primary</b>	<b>70.90<sub>9</sub></b>	<b>83.36<sub>8</sub></b>	<b>77.38<sub>10</sub></b>	<b>62.48<sub>7</sub></b>	<b>62.53<sub>5</sub></b>	<b>62.50<sub>7</sub></b>	<b>69.51<sub>6</sub></b>
	<b>10 PMI-cool-primary</b>	<b>68.79<sub>10</sub></b>	<b>79.94<sub>10</sub></b>	<b>80.00<sub>9</sub></b>	<b>47.81<sub>12</sub></b>	<b>70.58<sub>2</sub></b>	<b>57.00<sub>9</sub></b>	<b>56.73<sub>12</sub></b>
<b>11</b>	UH-PRHLT-contrastive1	67.57	79.50	77.08	54.10	50.11	52.03	62.45
	<b>11 UH-PRHLT-primary</b>	<b>67.42<sub>11</sub></b>	<b>79.38<sub>11</sub></b>	<b>76.97<sub>11</sub></b>	<b>55.64<sub>10</sub></b>	<b>46.80<sub>11</sub></b>	<b>50.84<sub>11</sub></b>	<b>63.21<sub>10</sub></b>
	UH-PRHLT-contrastive2	67.33	79.34	76.73	54.97	49.13	51.89	62.97
<b>12</b>	<b>QAIIT-primary</b>	<b>62.24<sub>12</sub></b>	<b>75.41<sub>12</sub></b>	<b>70.58<sub>12</sub></b>	<b>50.28<sub>11</sub></b>	<b>53.50<sub>9</sub></b>	<b>51.84<sub>10</sub></b>	<b>59.60<sub>11</sub></b>
	QAIIT-contrastive2	61.93	75.22	69.95	49.48	49.96	49.72	58.93
	QAIIT-contrastive1	61.80	75.12	69.76	49.85	50.94	50.39	59.24
	Baseline 1 (IR)	<b>59.53</b>	<b>72.60</b>	<b>67.83</b>	—	—	—	—
	Baseline 2 (random)	52.80	66.52	58.71	40.56	74.57	52.55	45.26
	Baseline 3 (all 'true')	—	—	—	40.64	100.00	<b>57.80</b>	40.64
	Baseline 4 (all 'false')	—	—	—	—	—	—	<b>59.36</b>

Table 1: **Subtask A, English (Question-Comment Similarity):** results for all submissions. The first column shows the rank of the primary runs with respect to the official MAP score. The second column contains the team’s name and its submission type (primary vs. contrastive). The following columns show the results for the primary, and then for other, unofficial evaluation measures. The subindices show the rank of the

# Natural Language Parsing tool: RevNLT

UK Economy News Headlines - FT.com - Mozilla Firefox

File Modifica Visualizza Cronologia Segnalibri Strumenti Aiuto

http://www.ft.com/world/uk/economy

Più visitati Corso: Basi di dati Gruppi Posta :: Benvenuto a H... ClustrMaps - map of vi... UniversitaCedol Tree Kernels in SVM-lig... Net RicercaAteneo Keysrc Calls EMEROTECA GEMS2010

The image displays a natural language parsing tree for the sentence: "Mortgage approvals fell sharply in June, lending yet more weight to the theory...". The tree is rooted at the top with a node labeled "1.00 V\_PP". This node branches into three children: "1.00 V\_Sog", "1.00 V\_Adv", and "1.00 Adv\_PP".

- The "1.00 V\_Sog" node branches into "Mortgage\_approvals" (type Nom) and "fell" (type VerFin, labeled "Sentence").
- The "1.00 V\_Adv" node branches into "sharply" (type Adv).
- The "1.00 Adv\_PP" node branches into "in\_June" (type Prep) and "lending\_yet\_more\_weight" (type Nom).
- The "in\_June" node branches into "in" (type IN, morph invariante) and "'June'" (type NNP, morph mas.fem.sing.plur.).
- The "lending\_yet\_more\_weight" node branches into "lending" (type NN, morph mas.fem.sing.), "yet" (type RB, morph invariante), and "more" (type JJR, morph mas.fem.plur.sing.).
- The "lending\_yet\_more\_weight" node also branches into "to\_the\_theory" (type Prep) and another node (partially visible, type NP\_PP).

Business Education  
Personal Finance  
Arts & Leisure  
Wealth  
In depth

Britain's place in the world, and how far it has travelled since 1947 - Jul-29

**Gilts lose lustre for overseas investors**  
Flight from eurozone risk to UK government bonds is moderating - Jul-29

with Alex Barker and Jim Pickard

Mechanical & Electrical Engineering  
Deputy Director of Finance  
London Ambulance Service  
RECRUITERS

http://www.ft.com/westminster

Italiano (Italia)

Today is: 2006-07-06 17:14:55

00:00:02:39

Timeline bar with play, stop, and volume controls.

Info	Transcription	Semantic Analysis	Content Analysis
00:06:36	chamonix <b>america</b> dove perde forza ma fa sempre paura l' uragano di mallarme <b>italia</b> andiamo nel centro		
00:06:41	che in florida riguardasse cento km di costa sull' atlantico si e' formata nel frattempo un' altra tempesta tropicale		
00:06:52	ha lasciato una riviera messicana dello jucker puntando verso la florida l' uragano delle corde <b>wilma</b> il dodicesimo ciclone di una stagione eco dell' atmosfera piu' di qualcuno lavatrici su strada a festeggiare lo scampato pericolo mentre dall' altra l' emergenza ha segnato l' inizio dei sa scarseggiano cibo e acqua si e' costretti a fare i conti con la sopravvivenza ad attraversare queste strade inondate sferzata dal		
00:07:22	vento la pioggia per raggiungere i centri della croce rossa vengono distribuiti ieri alla popolazione <b>dino risi</b> ma ha lasciato otto vittime soltanto migliaia di casi devastato la rete ospedaliera abbattuto centrali elettriche che ha causato danni a un milione di persone in florida e' attesa per		
00:07:44	e nelle isole di <b>kiss</b> e' gia' iniziata la grande fuga non bastasse sull' atlantico a sud di porto rico si e' formata falla venti di <b>hemingway</b> le		
00:07:52	nessuna tempesta tropicale della stagione la buona notizia che dovrebbe essere innocua la brutta notizia che la stagione degli uragani		
00:07:59	non e' ancora finita nulla fino al trentanove e c' e' stato una sciagura		
00:08:05	in nigeria		

- TG1 - 2005-10-23
- Other Classification
- Other Classification
- Other Classification
- Ambiente, Natura e Territorio
- Ambiente, Natura e Territorio**
- Other Classification
- Politica, Partiti, Istituzioni e Sindacati
- Politica, Partiti, Istituzioni e Sindacati
- Other Classification
- Other Classification
- Other Classification
- Other Classification
- Usi e costumi
- Other Classification
- Sanita' e Salute
- Giustizia, Criminalita' e Sicurezza
- Other Classification
- Other Classification
- Giustizia, Criminalita' e Sicurezza
- Other Classification
- Other Classification
- Musica e Spettacolo
- Sport
- Sport

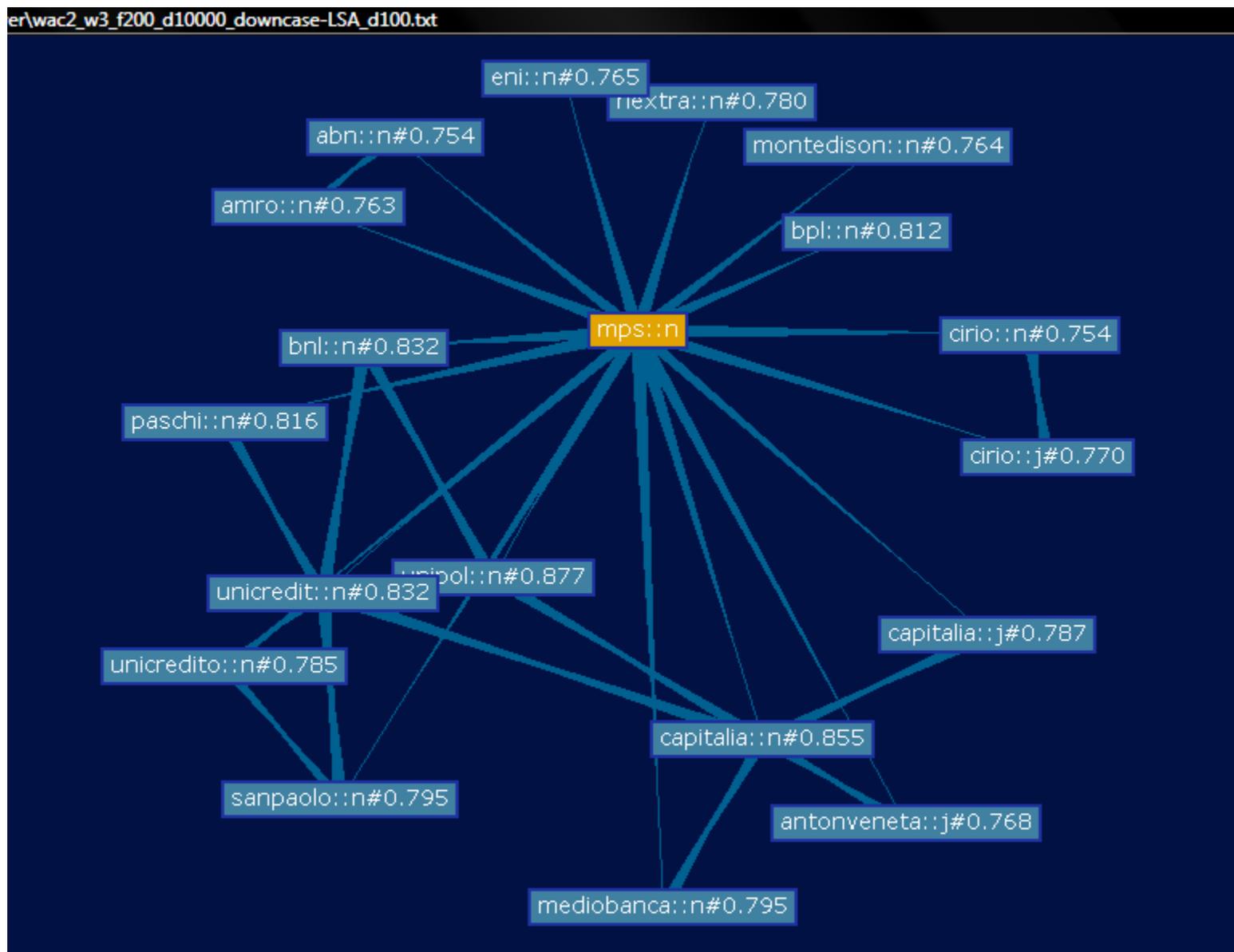
TIMELINE

00:06:36:10 00:06:51:07 00:06:54:20 00:06:56:01 00:06:57:10 00:07:00:08

00:07:08:18 00:07:12:02 00:07:15:23 00:07:19:13 00:07:22:02 00:07:26:06

00:07:37:08 00:07:39:15 00:07:45:20 00:07:55:06 00:07:57:07 00:07:59:13

# Vector Spaces for Lexical Semantics



# Sperimentazioni attive su

<http://mscoco.org/>



cocodataset@outlook.com

Home People Explore **Dataset** External

MS COCO image and annotations



a group of people point towards a green bus.

COCO-Text annotations

- 1) 'men',  
legible  
machine printed  
English
- 2) 'transport',  
legible  
machine printed  
English

OCR results:  
CORRECT

## COCO-Text

COCO-Text is for both text detection and recognition. The dataset annotates scene text with transcriptions along with attributes such as legibility, printed or handwritten text.



## FM-IQA

The Freestyle Multilingual Image Question Answering (FM-IQA) dataset contains over 120,000 images and 250,000 freestyle Chinese question-answer pairs and their English translations.



What color are her eyes?  
What is the mustache made of?



How many slices of pizza are there?  
Is this a vegetarian pizza?



Is this person expecting company?  
What is laid under the tree?



Does it appear to be rainy?  
Does this person have 20/20 vision?

## VQA

VQA is a new dataset containing open-ended questions about images. These questions require an understanding of vision, language and commonsense knowledge to answer.

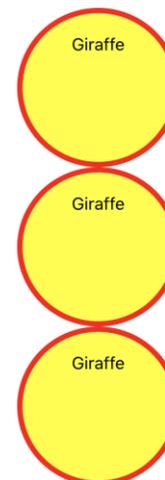
# MS GPT-3

- <https://twitter.com/i/status/1282676454690451457>

## Describe a layout.

Just describe any layout you want, and it'll try to render below!

```
<div style={{padding: 10}}> <div style={{backgroundColor: 'yellow', border: '5px solid red', borderRadius: '50%', padding: 20, width: 100, height: 100}}>Giraffe</div> <div style={{backgroundColor: 'yellow', border: '5px solid red', borderRadius: '50%', padding: 20, width: 100, height: 100}}>Giraffe</div> <div style={{backgroundColor: 'yellow', border: '5px solid red', borderRadius: '50%', padding: 20, width: 100, height: 100}}>Giraffe</div> </div>
```



# References

- Mitchell, Tom. M. 1997. *Machine Learning*. New York: McGraw-Hill.
- [Kernel machines, neural networks and graphical models](#), P. Frasconi, A. Sperduti, A. Starita, Rivista AI\*IA Numero speciale per i “50 anni di IA”, 2007.
- Very good video lectures by Andrew Ng (Stanford) <http://academicearth.org/courses/machine-learning>