CORSO DI *WEB MINING E RETRIEVAL*
- INTRODUZIONE AL DEEP LEARNING -

Corso di Laurea in Informatica, Ing. Gestionale, Ing. Internet, Ing. Informatica,

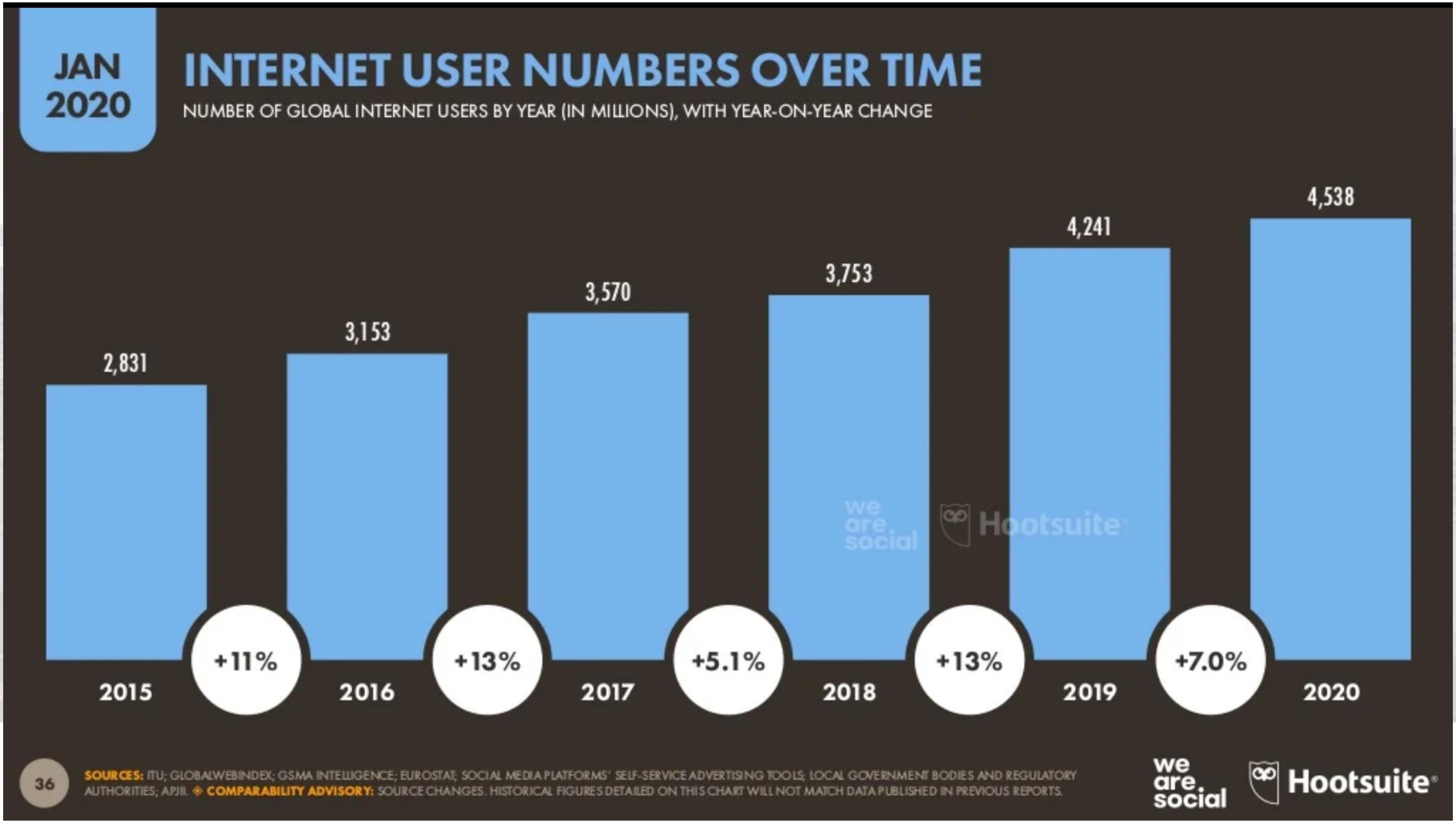
(a.a. 2022-2023)

Roberto Basili

Overview

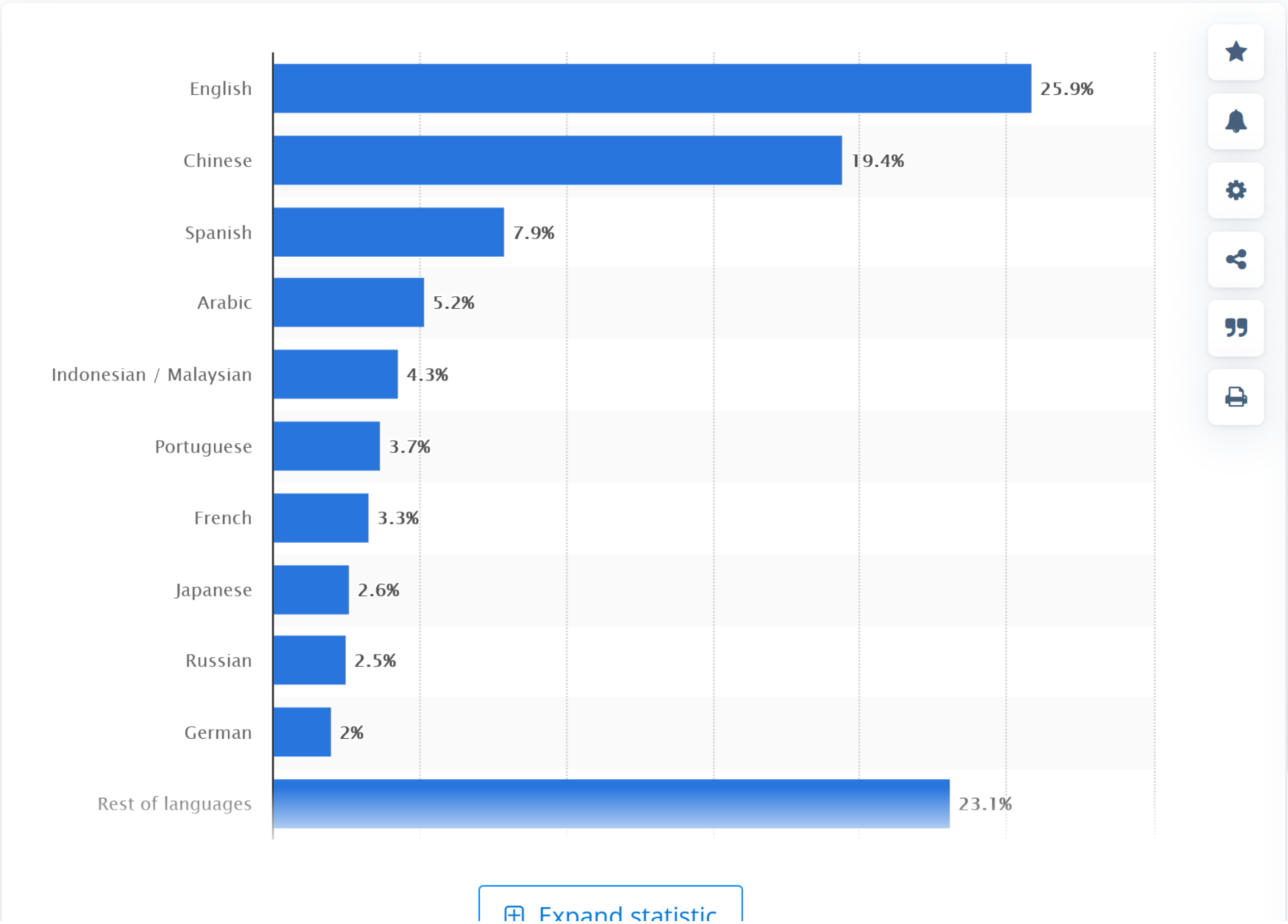
- Web Mining & Retrieval: Motivations & perspectives
  - Web, User-generated contents, Social Media
  - The role of *learning*
  - What is Machine Learning?
  - Data-driven algorithms: sources of complexity
- Main Applications
  - Intelligent Web Search
  - User Profiling for Marketing or Brand reputation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

Internet statistics (Jan 2021)

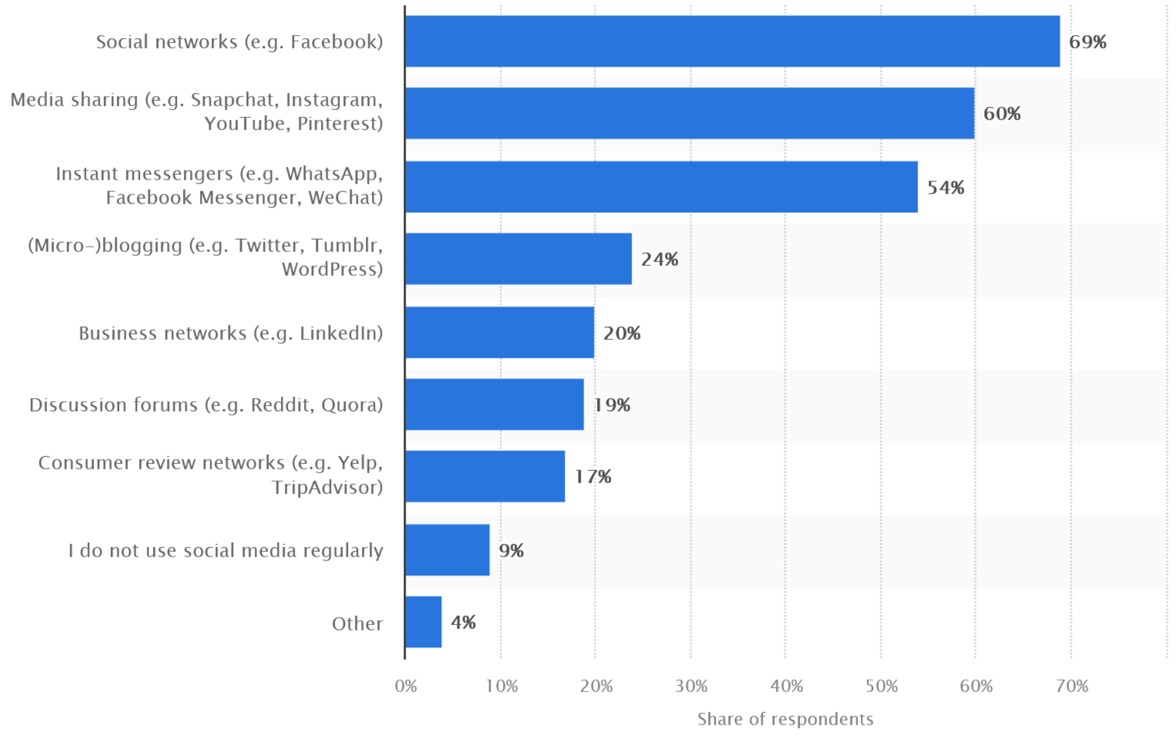


Internet Statistics (Jan 2020)

Most common languages used on the internet as of January 2020, by share of internet users



What kinds of social media do you use regularly?

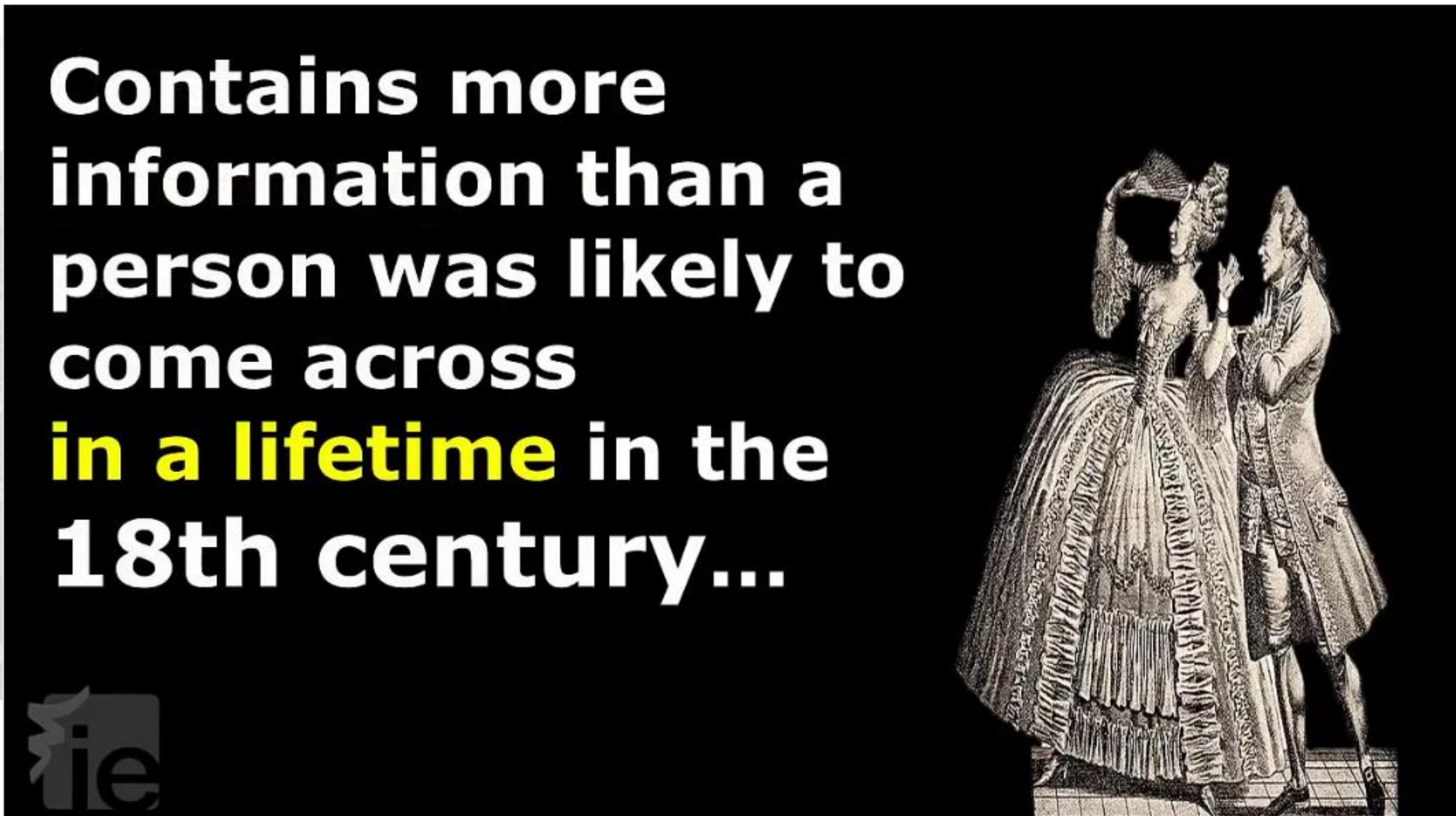


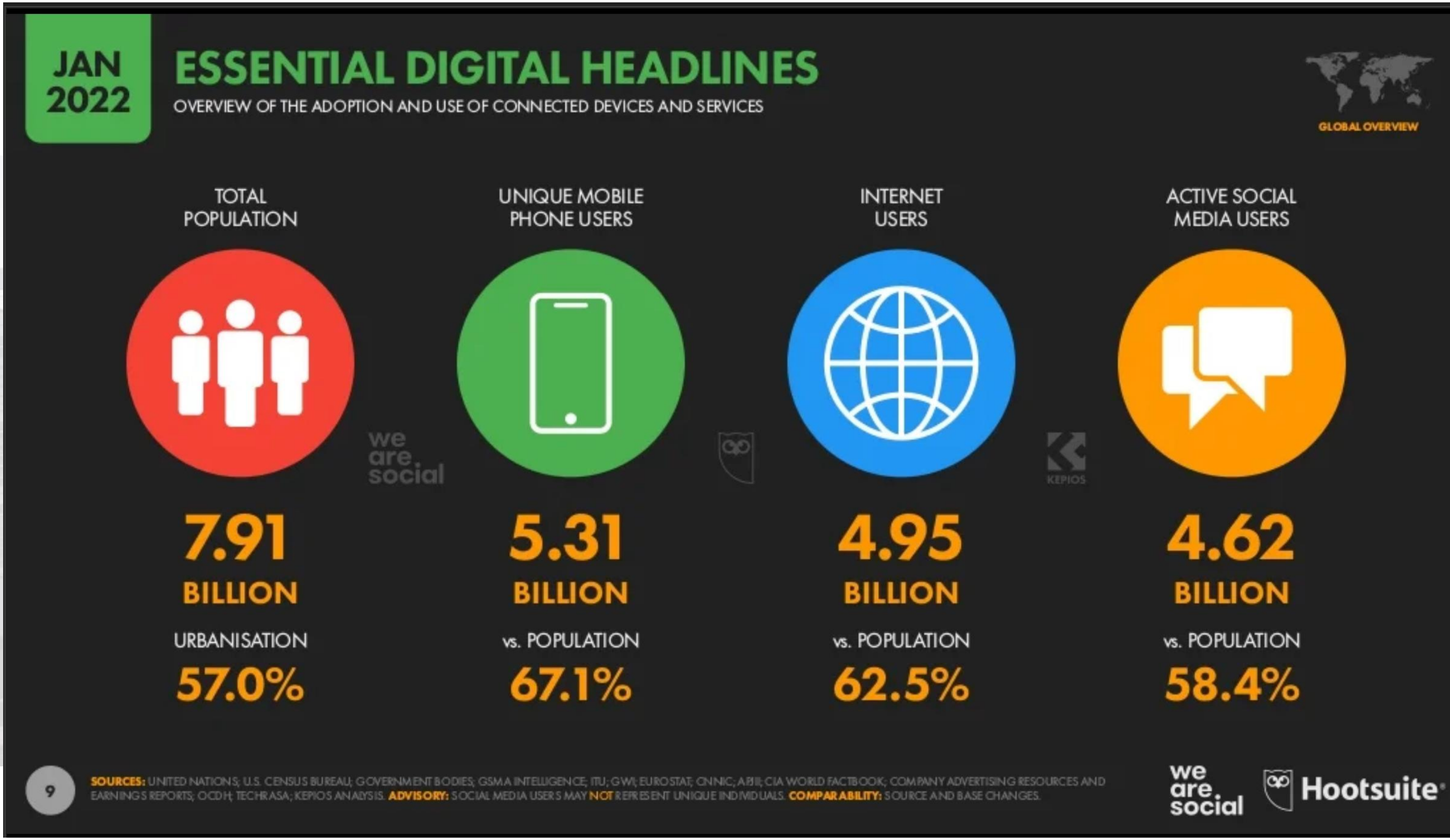
Details:

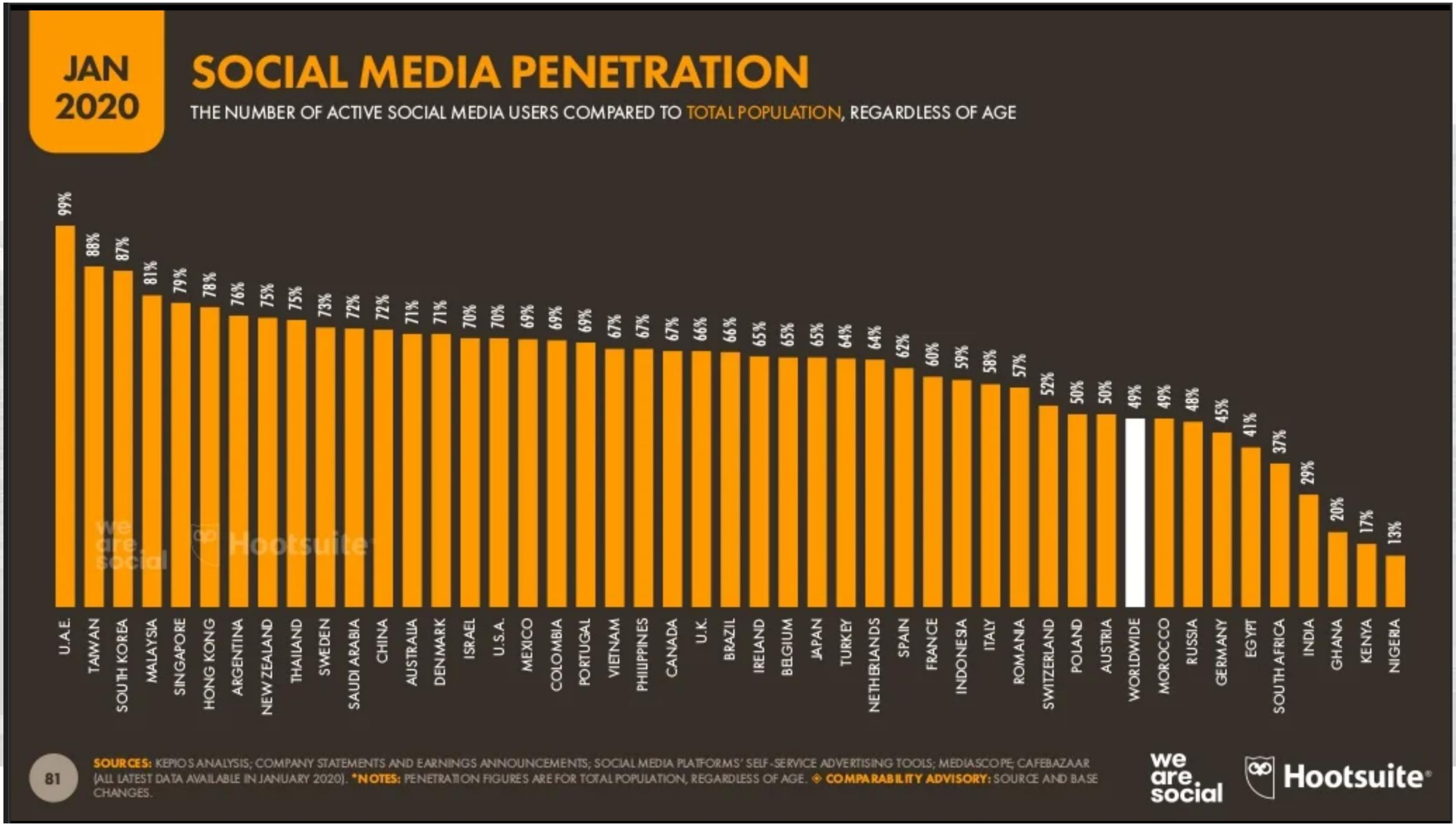
Show source

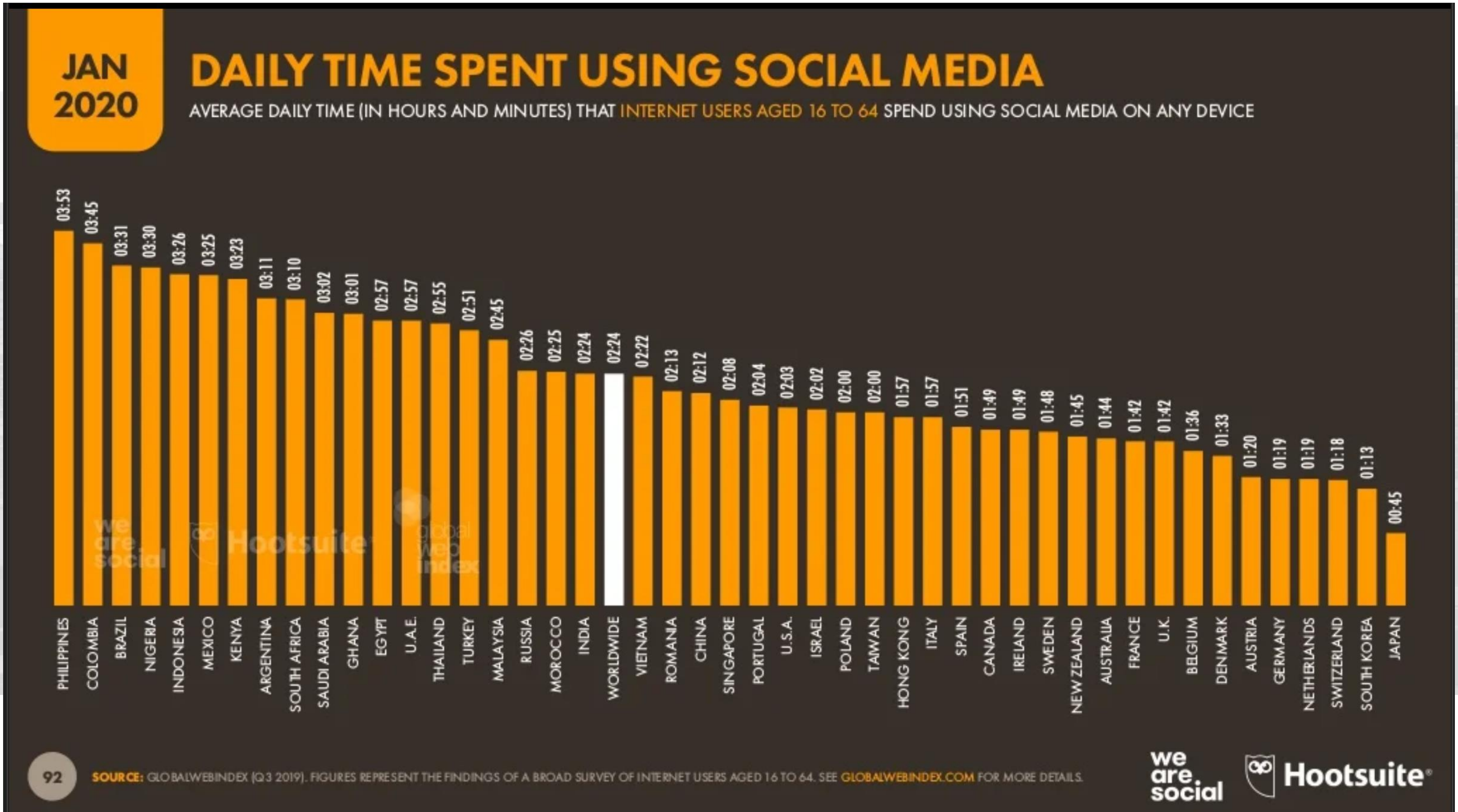
Do you know

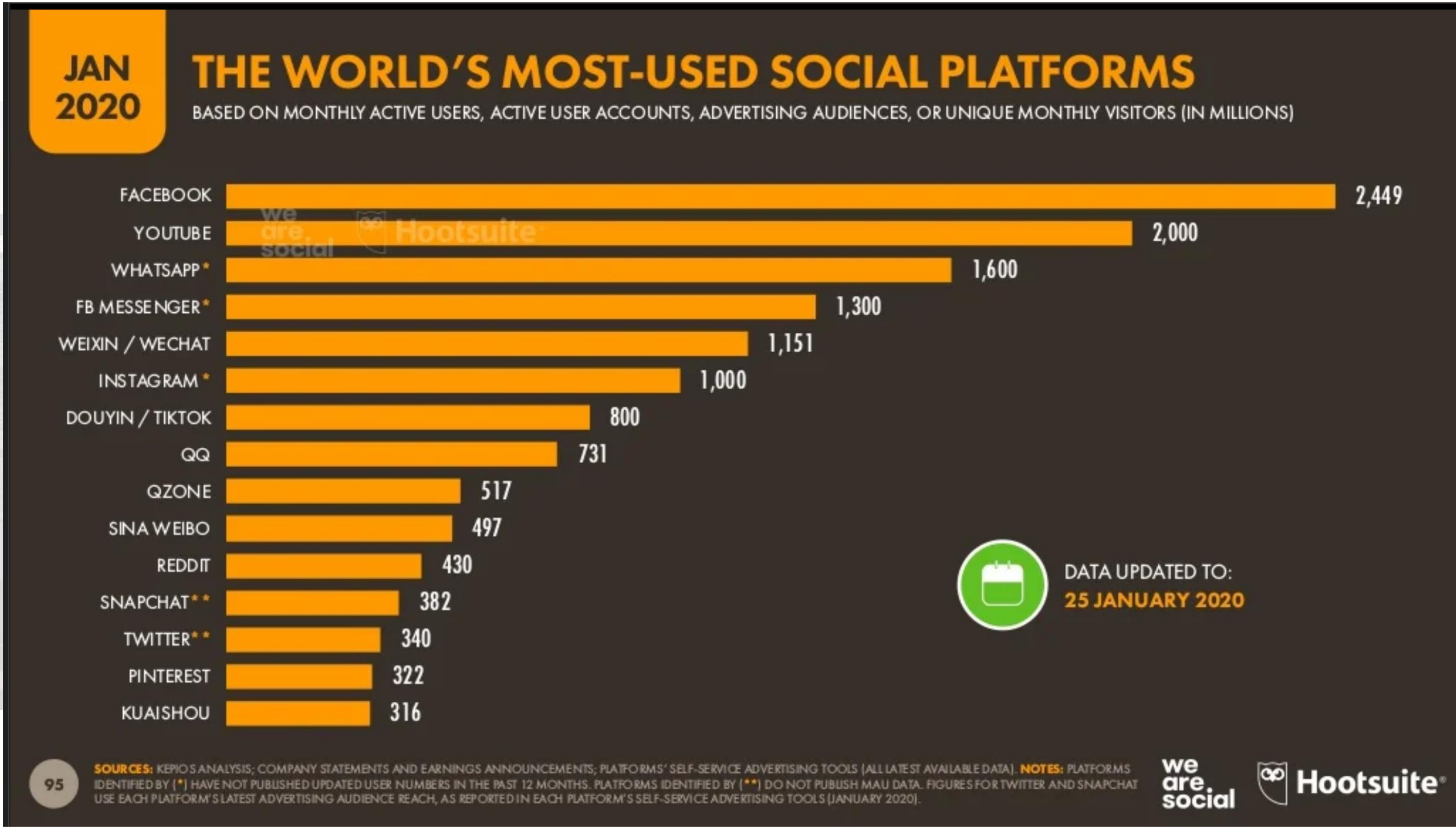
Do you know

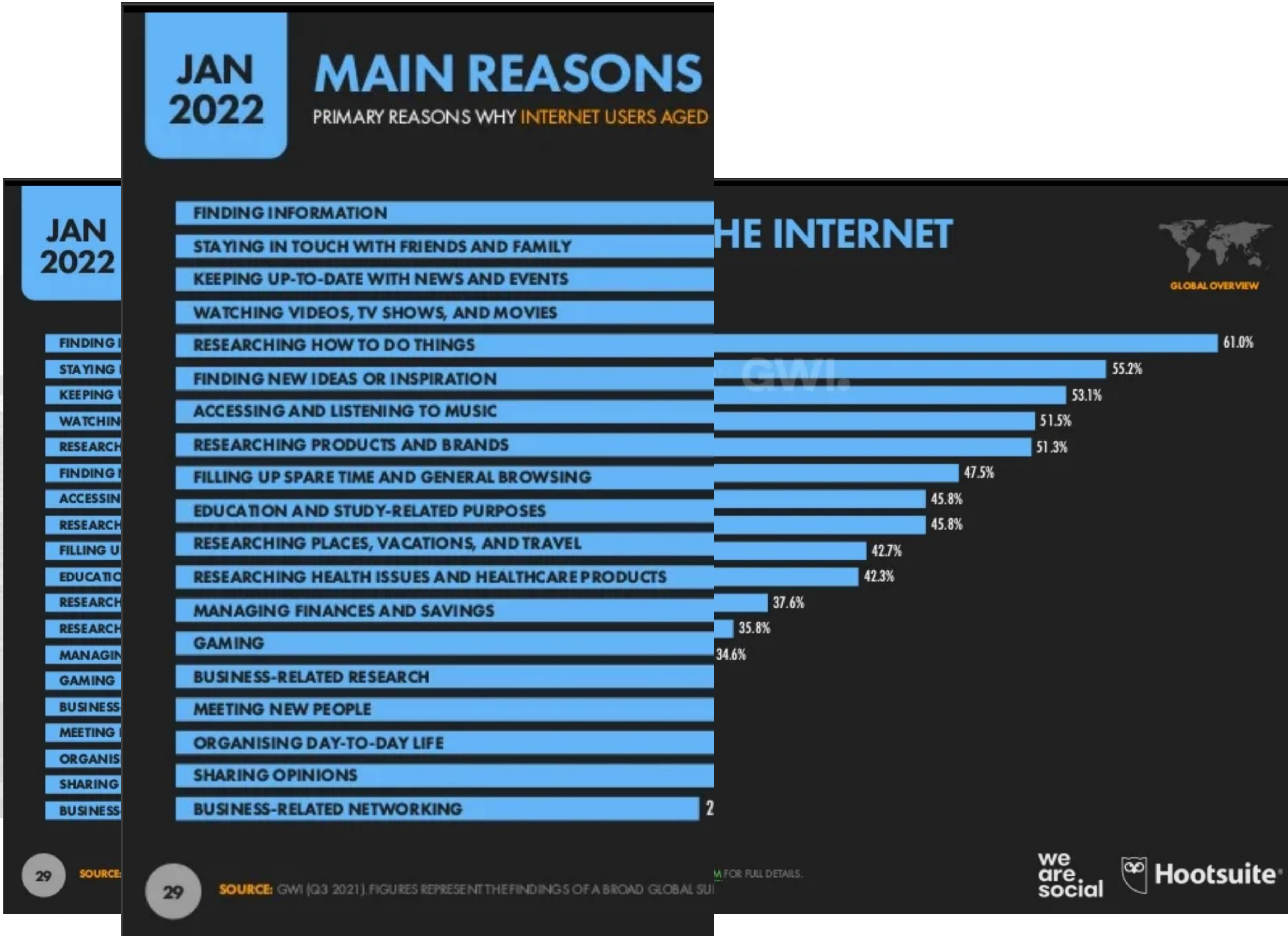

Contains more
information than a
person was likely to
come across
in a lifetime in the
18th century...


JAN 2022
ESSENTIAL DIGITAL HEADLINES
OVERVIEW OF THE ADOPTION AND USE OF CONNECTED DEVICES AND SERVICES
GLOBAL OVERVIEW
TOTAL POPULATION
7.91
BILLION
URBANISATION
57.0%
UNIQUE MOBILE PHONE USERS
5.31
BILLION
vs. POPULATION
67.1%
INTERNET USERS
4.95
BILLION
vs. POPULATION
62.5%
ACTIVE SOCIAL MEDIA USERS
4.62
BILLION
vs. POPULATION
58.4%
9
SOURCES: UNITED NATIONS; U.S. CENSUS BUREAU; GOVERNMENT BODIES; GSMA INTELLIGENCE; ITU; GWI; EUROSTAT; CNNIC; APJII; CIA WORLD FACTBOOK; COMPANY ADVERTISING RESOURCES AND EARNINGS REPORTS; OCDH; TECHRASA; KEPIOS ANALYSIS. ADVISORY: SOCIAL MEDIA USERS MAY NOT REPRESENT UNIQUE INDIVIDUALS. COMPARABILITY: SOURCE AND BASE CHANGES.
we are social
Hootsuite

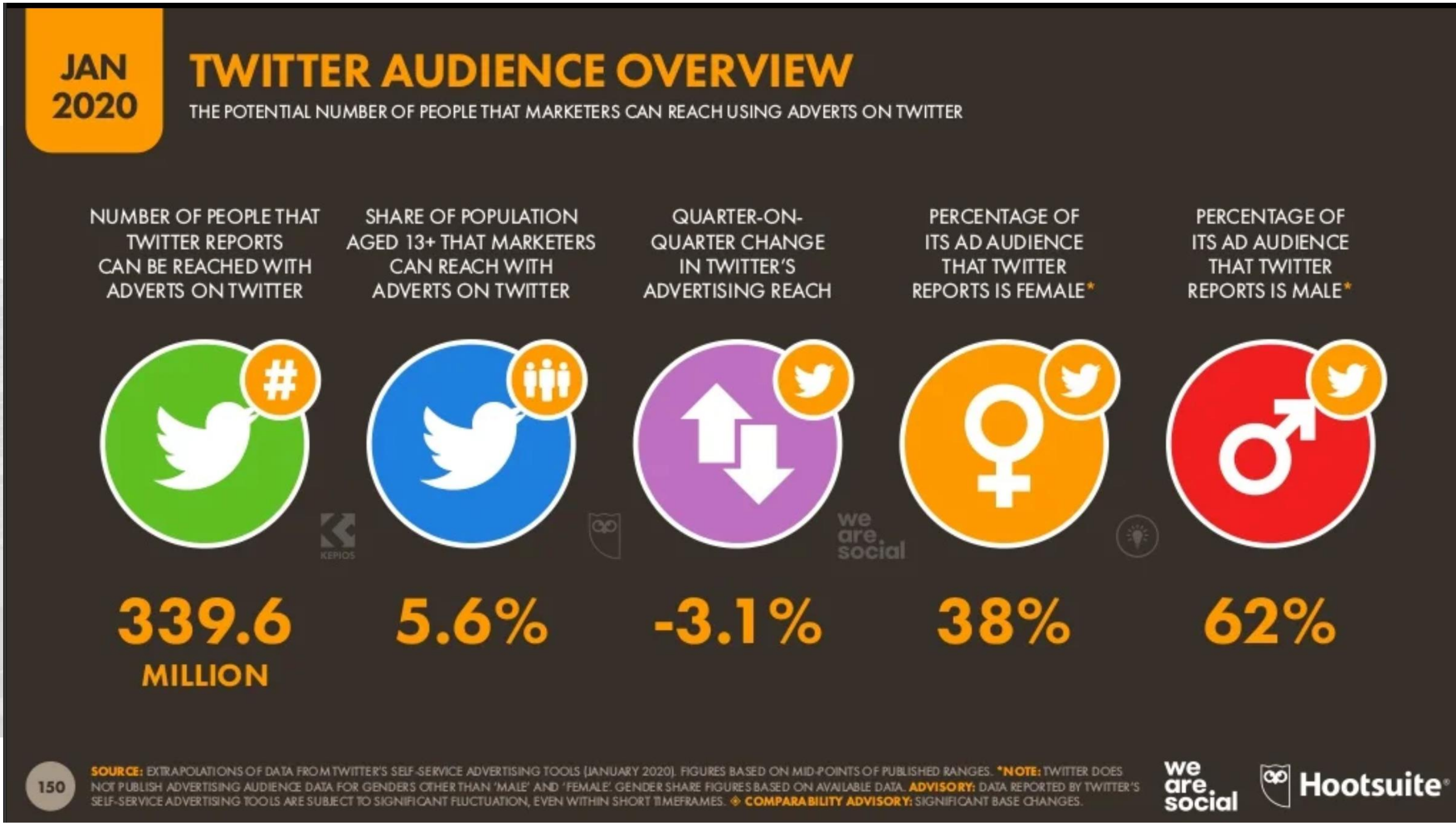

JAN 2020
SOCIAL MEDIA PENETRATION
THE NUMBER OF ACTIVE SOCIAL MEDIA USERS COMPARED TO TOTAL POPULATION, REGARDLESS OF AGE
U.A.E. 99%
TAIWAN 88%
SOUTH KOREA 87%
MALAYSIA 81%
SINGAPORE 79%
HONG KONG 78%
ARGENTINA 76%
NEW ZEALAND 75%
THAILAND 75%
SWEDEN 73%
SAUDI ARABIA 72%
CHINA 72%
AUSTRALIA 71%
DENMARK 71%
ISRAEL 70%
U.S.A. 70%
MEXICO 69%
COLOMBIA 69%
PORTUGAL 69%
VIETNAM 67%
PHILIPPINES 67%
CANADA 67%
U.K. 66%
BRAZIL 66%
IRELAND 65%
BELGIUM 65%
JAPAN 65%
TURKEY 64%
NETHERLANDS 64%
SPAIN 62%
FRANCE 60%
INDONESIA 59%
ITALY 58%
ROMANIA 57%
SWITZERLAND 52%
POLAND 50%
AUSTRIA 50%
WORLDWIDE 49%
MOROCCO 49%
RUSSIA 48%
GERMANY 45%
EGYPT 41%
SOUTH AFRICA 37%
INDIA 29%
GHANA 20%
KENYA 17%
NIGERIA 13%
81
SOURCES: KEPIOS ANALYSIS; COMPANY STATEMENTS AND EARNINGS ANNOUNCEMENTS; SOCIAL MEDIA PLATFORMS' SELF-SERVICE ADVERTISING TOOLS; MEDIASCOPE; CAFEBAZAAR (ALL LATEST DATA AVAILABLE IN JANUARY 2020). *NOTES: PENETRATION FIGURES ARE FOR TOTAL POPULATION, REGARDLESS OF AGE. COMPARABILITY ADVISORY: SOURCE AND BASE CHANGES.
we are social
Hootsuite


JAN 2020
DAILY TIME SPENT USING SOCIAL MEDIA
AVERAGE DAILY TIME (IN HOURS AND MINUTES) THAT INTERNET USERS AGED 16 TO 64 SPEND USING SOCIAL MEDIA ON ANY DEVICE
03:53 PHILIPPINES
03:45 COLOMBIA
03:31 BRAZIL
03:30 NIGERIA
03:26 INDONESIA
03:25 MEXICO
03:23 KENYA
03:11 ARGENTINA
03:10 SOUTH AFRICA
03:02 SAUDI ARABIA
03:01 GHANA
02:57 EGYPT
02:57 U.A.E.
02:55 THAILAND
02:51 TURKEY
02:45 MALAYSIA
02:26 RUSSIA
02:25 MOROCCO
02:24 INDIA
02:24 WORLDWIDE
02:22 VIETNAM
02:13 ROMANIA
02:12 CHINA
02:08 SINGAPORE
02:04 PORTUGAL
02:03 U.S.A.
02:02 ISRAEL
02:00 POLAND
02:00 TAIWAN
01:57 HONG KONG
01:57 ITALY
01:51 SPAIN
01:49 CANADA
01:49 IRELAND
01:48 SWEDEN
01:45 NEW ZEALAND
01:44 AUSTRALIA
01:42 FRANCE
01:42 U.K.
01:36 BELGIUM
01:33 DENMARK
01:20 AUSTRIA
01:19 GERMANY
01:19 NETHERLANDS
01:18 SWITZERLAND
01:13 SOUTH KOREA
00:45 JAPAN
92
SOURCE: GLOBALWEBINDEX (Q3 2019). FIGURES REPRESENT THE FINDINGS OF A BROAD SURVEY OF INTERNET USERS AGED 16 TO 64. SEE GLOBALWEBINDEX.COM FOR MORE DETAILS.
we are social
Hootsuite


JAN 2020
THE WORLD'S MOST-USED SOCIAL PLATFORMS
BASED ON MONTHLY ACTIVE USERS, ACTIVE USER ACCOUNTS, ADVERTISING AUDIENCES, OR UNIQUE MONTHLY VISITORS (IN MILLIONS)
FACEBOOK 2,449
YOUTUBE 2,000
WHATSAPP* 1,600
FB MESSENGER* 1,300
WEIXIN / WECHAT 1,151
INSTAGRAM* 1,000
DOUYIN / TIKTOK 800
QQ 731
QZONE 517
SINA WEIBO 497
REDDIT 430
SNAPCHAT** 382
TWITTER** 340
PINTEREST 322
KUAISHOU 316
DATA UPDATED TO:
25 JANUARY 2020
95
SOURCES: KEPIOS ANALYSIS; COMPANY STATEMENTS AND EARNINGS ANNOUNCEMENTS; PLATFORMS' SELF-SERVICE ADVERTISING TOOLS (ALL LATEST AVAILABLE DATA). NOTES: PLATFORMS IDENTIFIED BY (*) HAVE NOT PUBLISHED UPDATED USER NUMBERS IN THE PAST 12 MONTHS. PLATFORMS IDENTIFIED BY (**) DO NOT PUBLISH MAU DATA. FIGURES FOR TWITTER AND SNAPCHAT USE EACH PLATFORM'S LATEST ADVERTISING AUDIENCE REACH, AS REPORTED IN EACH PLATFORM'S SELF-SERVICE ADVERTISING TOOLS (JANUARY 2020).
we are social
Hootsuite



JAN 2022
MAIN REASONS
PRIMARY REASONS WHY INTERNET USERS AGED
FINDING INFORMATION
STAYING IN TOUCH WITH FRIENDS AND FAMILY
KEEPING UP-TO-DATE WITH NEWS AND EVENTS
WATCHING VIDEOS, TV SHOWS, AND MOVIES
RESEARCHING HOW TO DO THINGS
FINDING NEW IDEAS OR INSPIRATION
ACCESSING AND LISTENING TO MUSIC
RESEARCHING PRODUCTS AND BRANDS
FILLING UP SPARE TIME AND GENERAL BROWSING
EDUCATION AND STUDY-RELATED PURPOSES
RESEARCHING PLACES, VACATIONS, AND TRAVEL
RESEARCHING HEALTH ISSUES AND HEALTHCARE PRODUCTS
MANAGING FINANCES AND SAVINGS
GAMING
BUSINESS-RELATED RESEARCH
MEETING NEW PEOPLE
ORGANISING DAY-TO-DAY LIFE
SHARING OPINIONS
BUSINESS-RELATED NETWORKING
29
SOURCE: GWI (Q3 2021). FIGURES REPRESENT THE FINDINGS OF A BROAD GLOBAL SU
JAN 2022
HE INTERNET
GLOBAL OVERVIEW
GWI.
61.0%
55.2%
53.1%
51.5%
51.3%
47.5%
45.8%
45.8%
42.7%
42.3%
37.6%
35.8%
34.6%
29
SOURCE:
FOR FULL DETAILS.
we are social
Hootsuite


JAN 2020
TWITTER AUDIENCE OVERVIEW
THE POTENTIAL NUMBER OF PEOPLE THAT MARKETERS CAN REACH USING ADVERTS ON TWITTER
NUMBER OF PEOPLE THAT TWITTER REPORTS CAN BE REACHED WITH ADVERTS ON TWITTER
SHARE OF POPULATION AGED 13+ THAT MARKETERS CAN REACH WITH ADVERTS ON TWITTER
QUARTER-ON-QUARTER CHANGE IN TWITTER'S ADVERTISING REACH
PERCENTAGE OF ITS AD AUDIENCE THAT TWITTER REPORTS IS FEMALE*
PERCENTAGE OF ITS AD AUDIENCE THAT TWITTER REPORTS IS MALE*
339.6 MILLION
5.6%
-3.1%
38%
62%
150
SOURCE: EXTRAPOLATIONS OF DATA FROM TWITTER'S SELF-SERVICE ADVERTISING TOOLS (JANUARY 2020). FIGURES BASED ON MID-POINTS OF PUBLISHED RANGES. *NOTE: TWITTER DOES NOT PUBLISH ADVERTISING AUDIENCE DATA FOR GENDERS OTHER THAN 'MALE' AND 'FEMALE'. GENDER SHARE FIGURES BASED ON AVAILABLE DATA. ADVISORY: DATA REPORTED BY TWITTER'S SELF-SERVICE ADVERTISING TOOLS ARE SUBJECT TO SIGNIFICANT FLUCTUATION, EVEN WITHIN SHORT TIMEFRAMES. ◈ COMPARABILITY ADVISORY: SIGNIFICANT BASE CHANGES.
we are social
Hootsuite

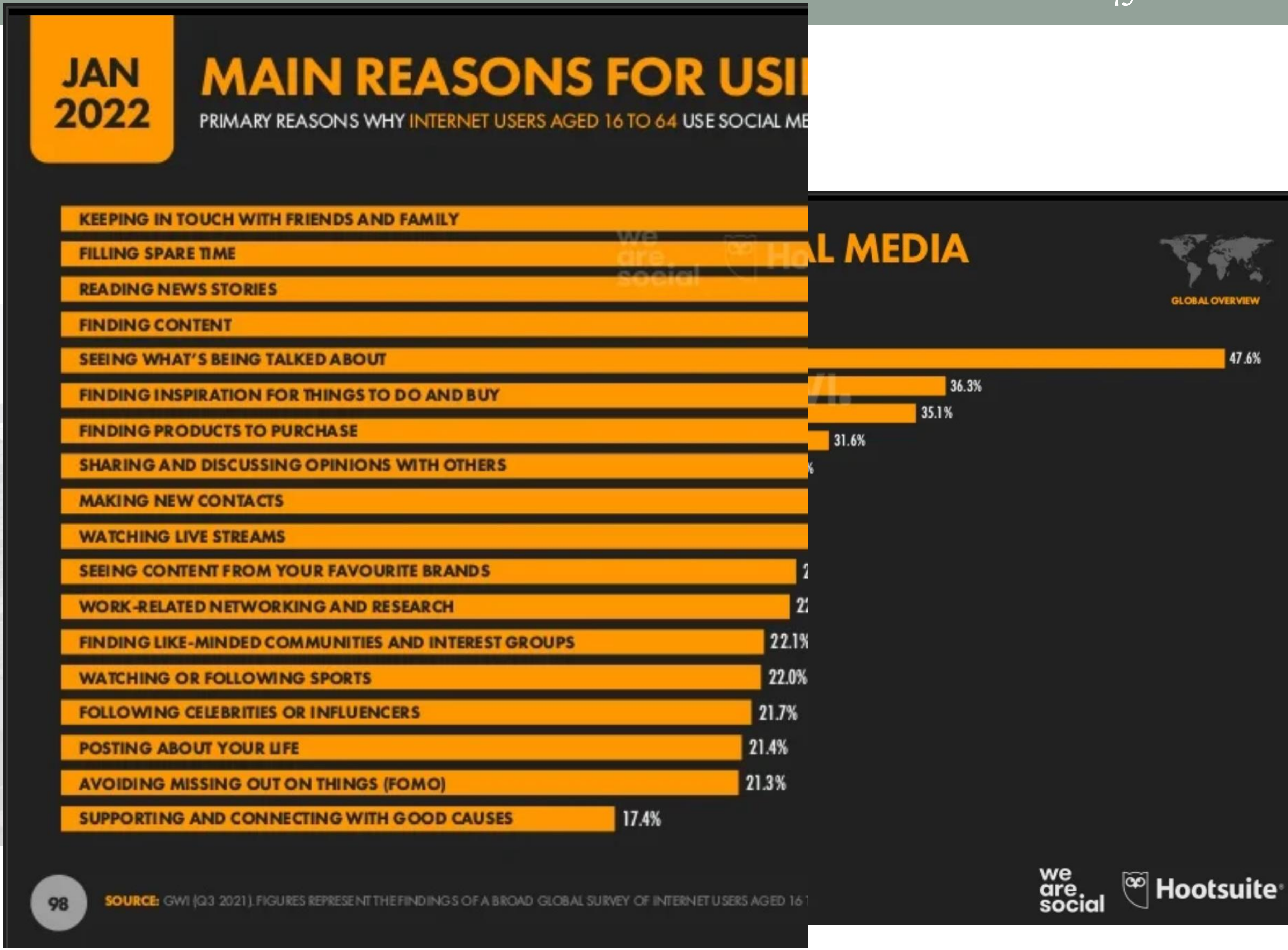
JAN 2020

MOST-USED EMOJI ON TWITTER

EMOJI THAT HAVE BEEN USED THE GREATEST NUMBER OF TIMES ON TWITTER (ALL TIME)

#	EMOJI	TIMES USED
01	😂	2,671,000,000
02	❤️	1,289,000,000
03	♻️	966,000,000
04	😍	964,000,000
05	😭	817,000,000
06	❤	743,000,000
07	😊	632,000,000
08	😒	500,000,000
09	💕	493,000,000
10	😘	475,000,000
11	😩	428,000,000
12	☺️	389,000,000
13	😔	382,000,000
14	😁	365,000,000
15	👌	359,000,000
16	😏	336,000,000
17	😉	309,000,000
18	👍	273,000,000
19	😅	258,000,000
20	🙏	246,000,000
21	👀	245,000,000
22	😌	238,000,000
23	😢	237,000,000
24	💔	236,000,000
25	😎	232,000,000
26	⬅️	229,000,000
27	🎶	217,000,000
28	💙	216,000,000
29	💜	212,000,000
30	✨	199,000,000
31	😳	198,000,000
32	🙌	193,000,000
33	🔥	191,000,000
34	💖	187,000,000
35	💯	182,000,000
36	🙈	181,000,000
37	✌️	168,000,000
38	😄	165,000,000
39	😪	163,000,000
40	😑	163,000,000

SOURCE: EMOJI TRACKER (JANUARY 2020). NOTE: VALUES HAVE BEEN ROUNDED TO THE NEAREST MILLION.



JAN 2022
MAIN REASONS FOR USI
L MEDIA
PRIMARY REASONS WHY INTERNET USERS AGED 16 TO 64 USE SOCIAL ME
GLOBAL OVERVIEW
KEEPING IN TOUCH WITH FRIENDS AND FAMILY
47.6%
FILLING SPARE TIME
36.3%
READING NEWS STORIES
35.1%
FINDING CONTENT
31.6%
SEEING WHAT'S BEING TALKED ABOUT
FINDING INSPIRATION FOR THINGS TO DO AND BUY
FINDING PRODUCTS TO PURCHASE
SHARING AND DISCUSSING OPINIONS WITH OTHERS
MAKING NEW CONTACTS
WATCHING LIVE STREAMS
SEEING CONTENT FROM YOUR FAVOURITE BRANDS
WORK-RELATED NETWORKING AND RESEARCH
FINDING LIKE-MINDED COMMUNITIES AND INTEREST GROUPS
22.1%
WATCHING OR FOLLOWING SPORTS
22.0%
FOLLOWING CELEBRITIES OR INFLUENCERS
21.7%
POSTING ABOUT YOUR LIFE
21.4%
AVOIDING MISSING OUT ON THINGS (FOMO)
21.3%
SUPPORTING AND CONNECTING WITH GOOD CAUSES
17.4%
98
SOURCE: GWI (Q3 2021). FIGURES REPRESENT THE FINDINGS OF A BROAD GLOBAL SURVEY OF INTERNET USERS AGED 16
we are social
Hootsuite

WE ARE SOCIAL'S PERSPECTIVE: SOCIAL IN 2020

SHIFTS IN HOW PEOPLE BEHAVE AND INTERACT ON SOCIAL

BAD INFLUENCE

Being a creator has lost its lo-fi sheen; many lifestyle influencers lead unrelatable lives, while celebrity 'creators' like Will Smith are blowing up on platforms like YouTube and TikTok. As a result, there's a growing backlash against influencer culture and the metrics that drive it.

In 2020, brands will look beyond likes, followers and reach to generate genuine engagement

ADDED VALUE

The internet has long been a wild west where intellectual property is barely there. But in a maturing digital frontier, creators have grown dedicated audiences who not only see value in their content, but recognise their style anywhere. As a result, communities are rallying to protect creators.

In 2020, brands will take greater steps to ensure they're being respectful of digital communities

RUNNING COMMENTARY

Audiences are increasingly willing to invest time and attention in content and narratives they deem to have a higher value. This isn't about a shift back to traditional media. It's about longer, more complex content designed to be consumed in-platform and on smaller screens.

In 2020, brands will tell more complex stories across multiple touchpoints on social

Dealing with *real* Social media data

WM&R: Motivations

- *What does Web Mining mean?*
- *Why Information Retrieval is involved?*
- *Why Machine Learning and mostly Deep Learning?*
- *Which are the contributions of IR/ML/NLP to technologies that support and exploit Web Data, Information and Knowledge?*
- *Which are the technological perspectives in the medium-long term?*

What is Web Mining?

- *Web Mining* refers to a body of technologies currently needed for the *exploitation of publicly available information from the Web and the IoT*
  - Contents: data but also … PEOPLE, LOCATIONS, EVENTS, CONCEPTS, TEMPORAL INFORMATION …
  - Relations:
    - Links within structured networks (retweets, follows, …)
    - Thematic, interpersonal and semantic associations
    - Similarities and Analogies among people, behaviours, preferences
  - On-Line Structured and semi-structured resources (e.g. Wikipedia)
  - Textual, Multimedia and Multilingual Contents
  - Trends e time-related information (community on-line behaviours)
  - Opinions, Preferences, Expectations

Why IR?

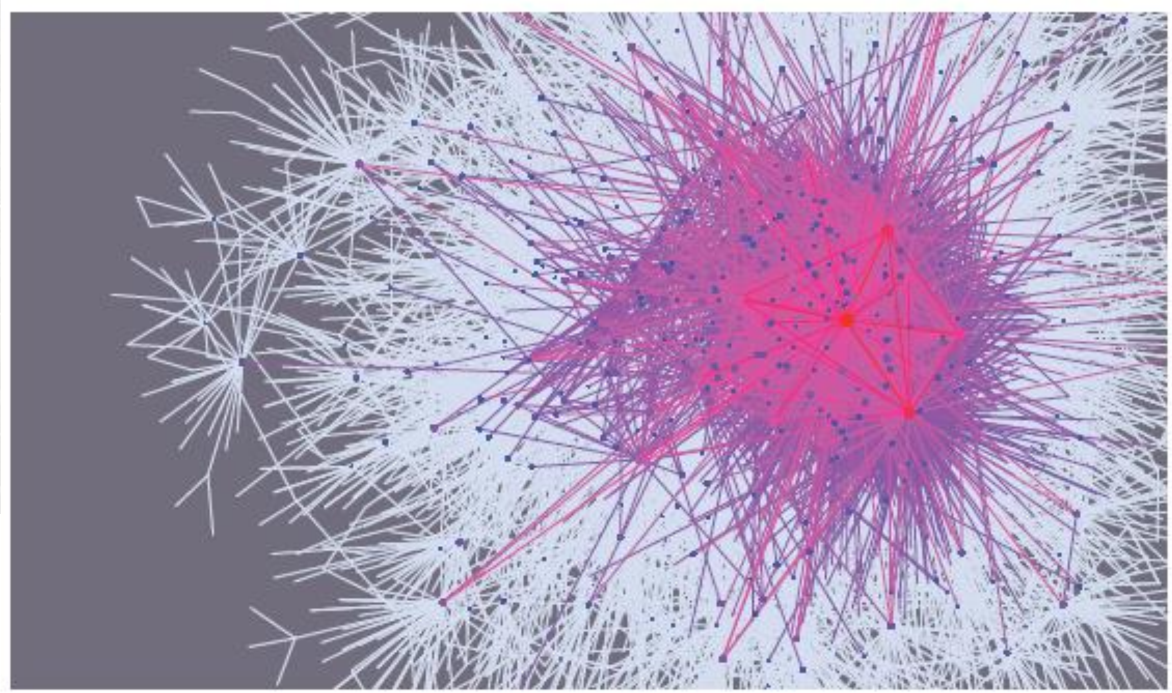
- The volumes involved in Web Mining pose the crucial problem of *locating information* beforehand
- Automatic information access is possible only if we solve the two major challenges
  - What is relevant
  - Where the relevant information is located
- Searching information corresponds to computing an uncertain function that models the *mapping* between information needs and the targeted data

Machine Learning vs IR?

- Web mining involve heterogeneous information that is characterized search as strongly uncertainy process
- The available information is characterized by:
  - Incompleteness:
    - Short queries as an incomplete description of the information need
  - Variability: Wealth of data vs. heterogeneity of formats and access modes
    - Contents are dispersed in various forms across data sources
  - Vague Requirements
    - Information is often implicit (i.e. partially and qualitatively expressed) in the operational contexts
  - Subjectivity
    - Relevance depends on the user and not just on the contents
  - Timeliness
  - Authority

Machine Learning vs. IR

- Uncertainty os so pervasive that exhaustive solutions (i.e. global *optima*) are not available or even not existing
- “*Finding diamonds in the rough*” (Fan Chung, UCSD)



Machine Learning vs. IR

- ML technologies offer a wide variety of algorithms, strategies and techniques for the induction of sub-optimal, but surprisingly effective, solutions from available data
- Through *learning* data can be effectively used to suggest retrieval hypothesis, that are models of the *mapping* function (Learning to search)
- What is the target of the learned function? To improve computational aspects of the currently applied processes, such as
    - Semantic Accuracy (i.e. best answers first)
    - System Responsiveness (i.e. reducing speed of the retrieval process)
    - Resource usage (i.e. more effective with less memory or input data)

Machine Learning

- Machine learning is **the study of computer algorithms that allow computer programs to automatically improve through experience.** (Tom Mitchell, *Machine Learning*, McGraw-Hill,1997)
- The evidence of the success of a learning process corresponds to the possibility of observing a measurable increment ΔP of performances in solving a task C on the basis of experiences *E* that the agent is able to gather during its lifecycle.
- The nature and complexity of the learning ability is fully confined to the ability of characterizing the primitive notions here involved:
  - Task *C*
  - Performance *P*
  - Experience *E*

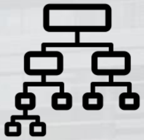
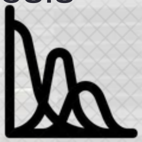
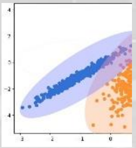
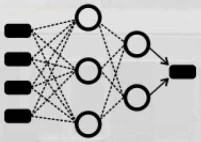
Experience and Learning

- Forms of experiences
- In chess games:
  - Data on previous matches, such as won challenges (or defeats) able to gather the utility (o inadequacy) of the strategies or moves there carried out.
  - Evaluation about individual moves offered by an external teacher (oracle, guide).
  - Adequacy of individual behaviour derived from self-observation, such as the capablilty of analyising matches against itself based on a existing explicit model of the rules and strategies of the game.

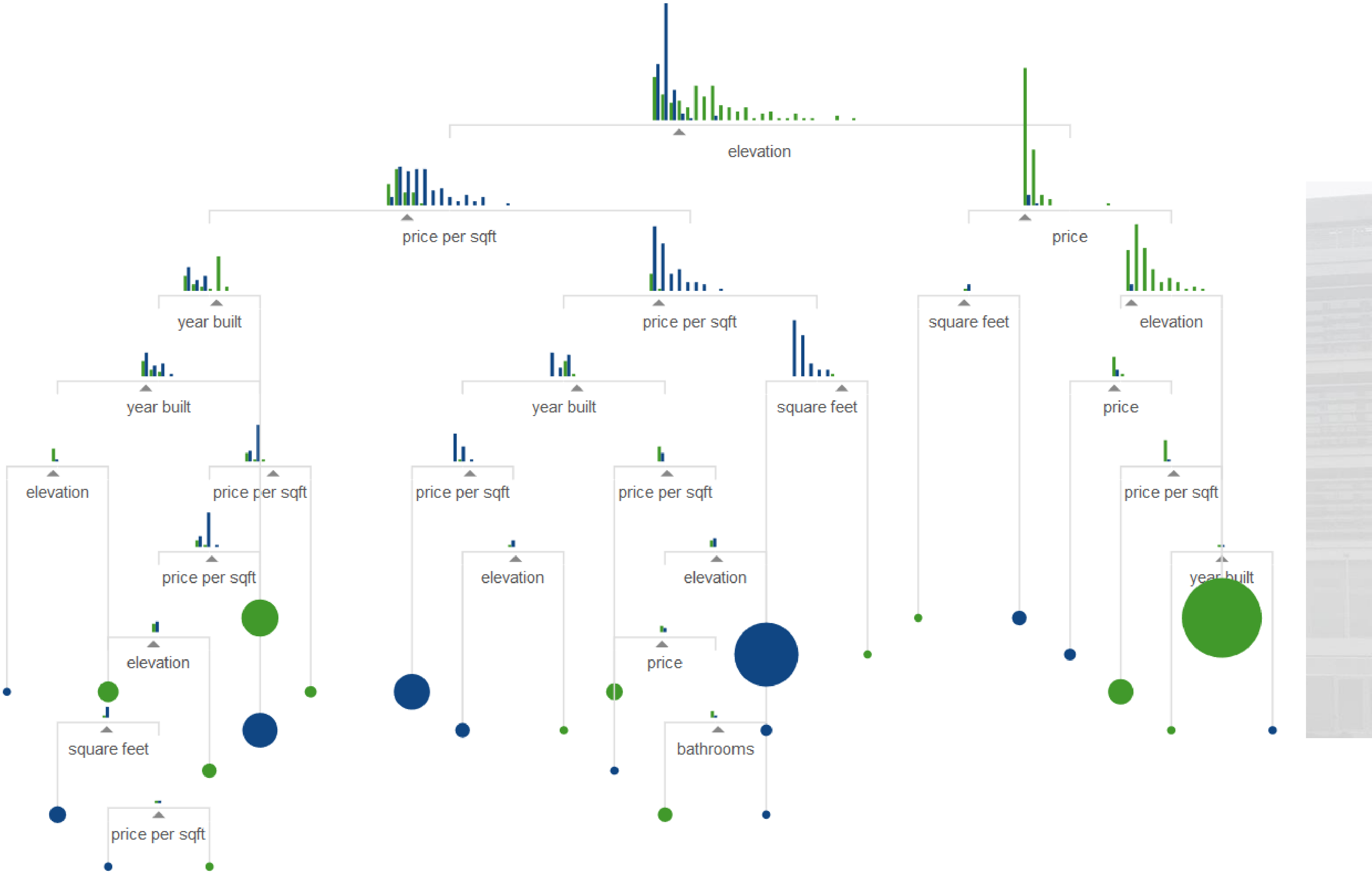
ML: a visual introduction

- See URL: http://www.r2d3.us/visual-intro-to-machine-learning-part-1/?imm_mid=0d76b4&cmp=em-data-na-na-newsltr_20150826

The mathemathics of Learning

- Learning corresponds to the induction of mathematical function (i.e. the decision rules) that may have a discrete as well as a continuous behaviour:
  
  - Logical functions, (ad es., decision trees)
    - Learning the rules that better explain the data
    - Induction: Recursive search for necessary and sufficient conditions.
  - Probabilistic Approaches:
    - Learning what is *most likely* to be the better decision, according to an hypothesis about the input distribution (e.g., Bayesian classification)
    
    - Induction: Estimate the Posterior Probability (as parameters of known laws).
  - Metric Approaches
    - Decision as discrimination in metric spaces (e.g. linear and non linear functions)
      
      
      - K-NN
      - Linear Classifiers, perceptrons, Neural Networks, Support Vector Machines,…
      - Modeling as vectorial embedding, spectral analysis (space transformations)
    - Induction: determine the optimal parametrization from specific function classes (e.g. multilayer networks, polynomials of degree *n*)

Es. Decision Tree Learning



Unsupervised Learning

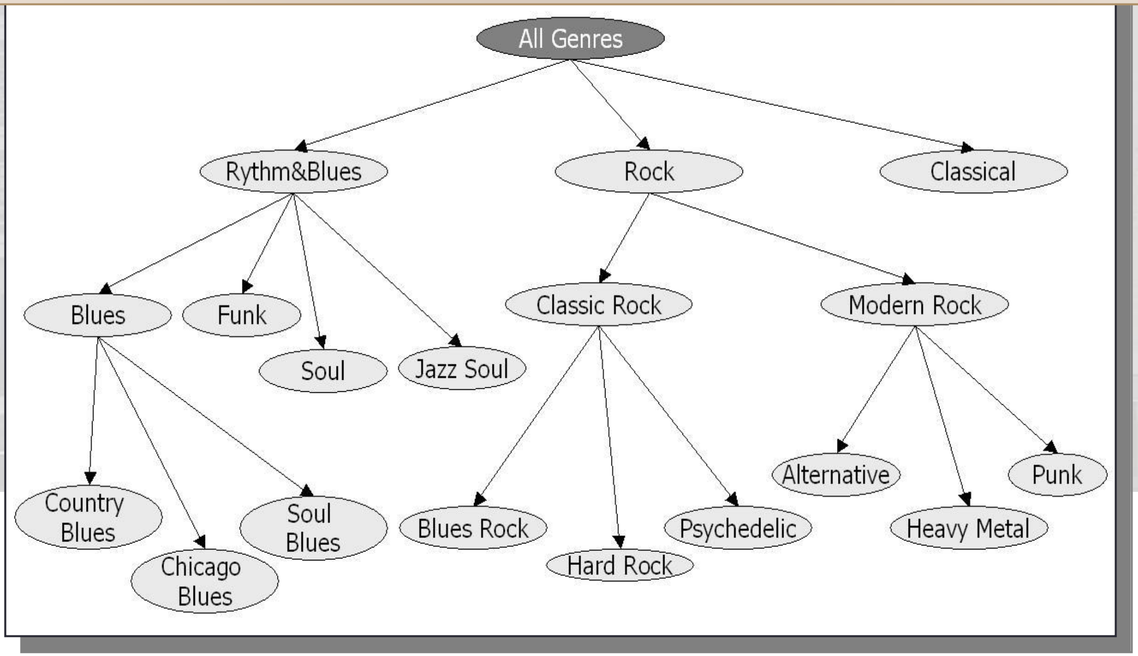
- When no oracle or knowledge of the task is available, learing may still be applied in several approaches:
  - Improve the current world model (*knowledge acquisition/discovery*)
  - Improve the efficiency of the currently available algorithms, through computational optimization
    - Better data structures representing the problem and the domain
    - Reduction of the processing steps required by the current models

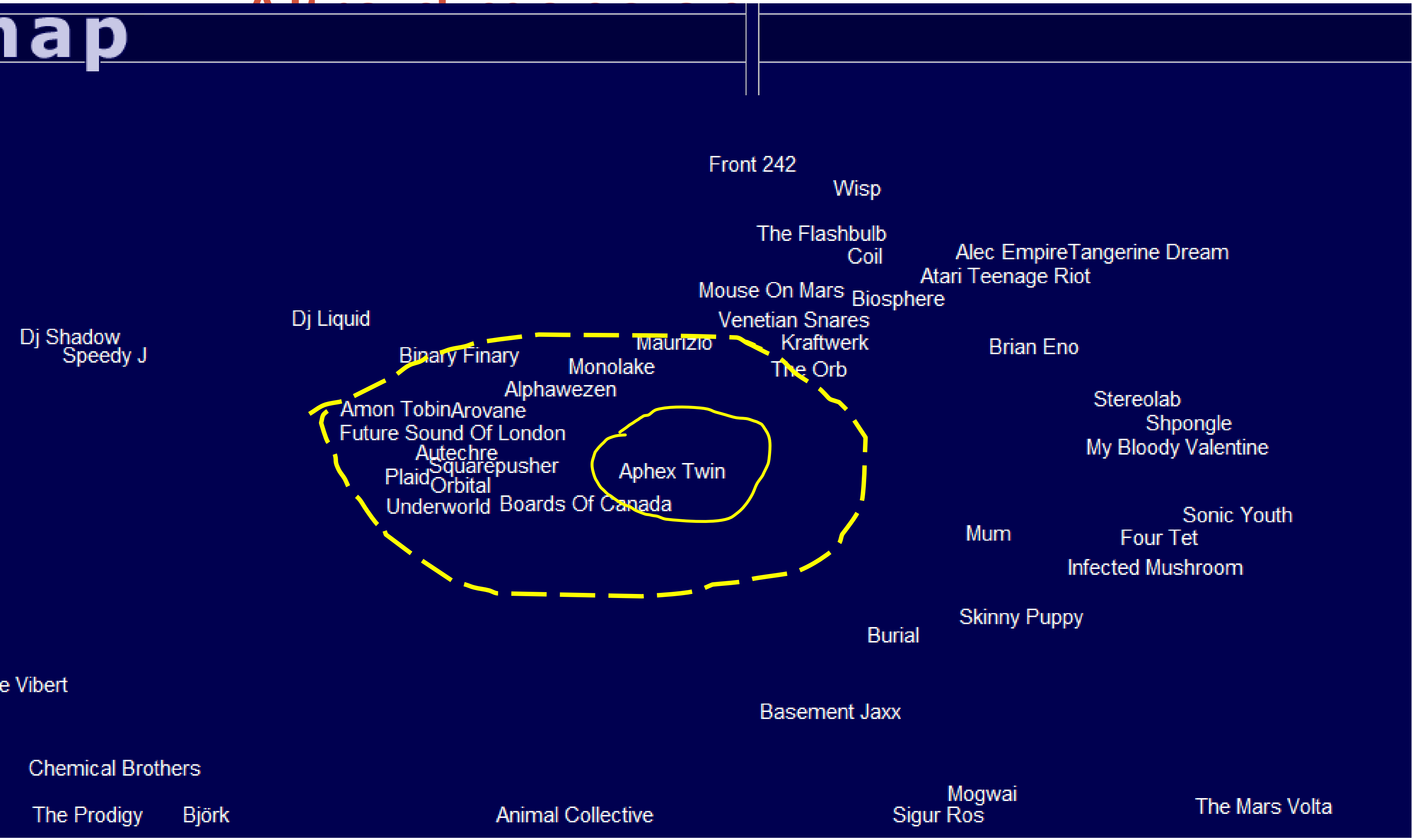
Unsupervised Learning

The induced hierachical model expresses a system of classes and relations able to imptove future interaction with the song collection

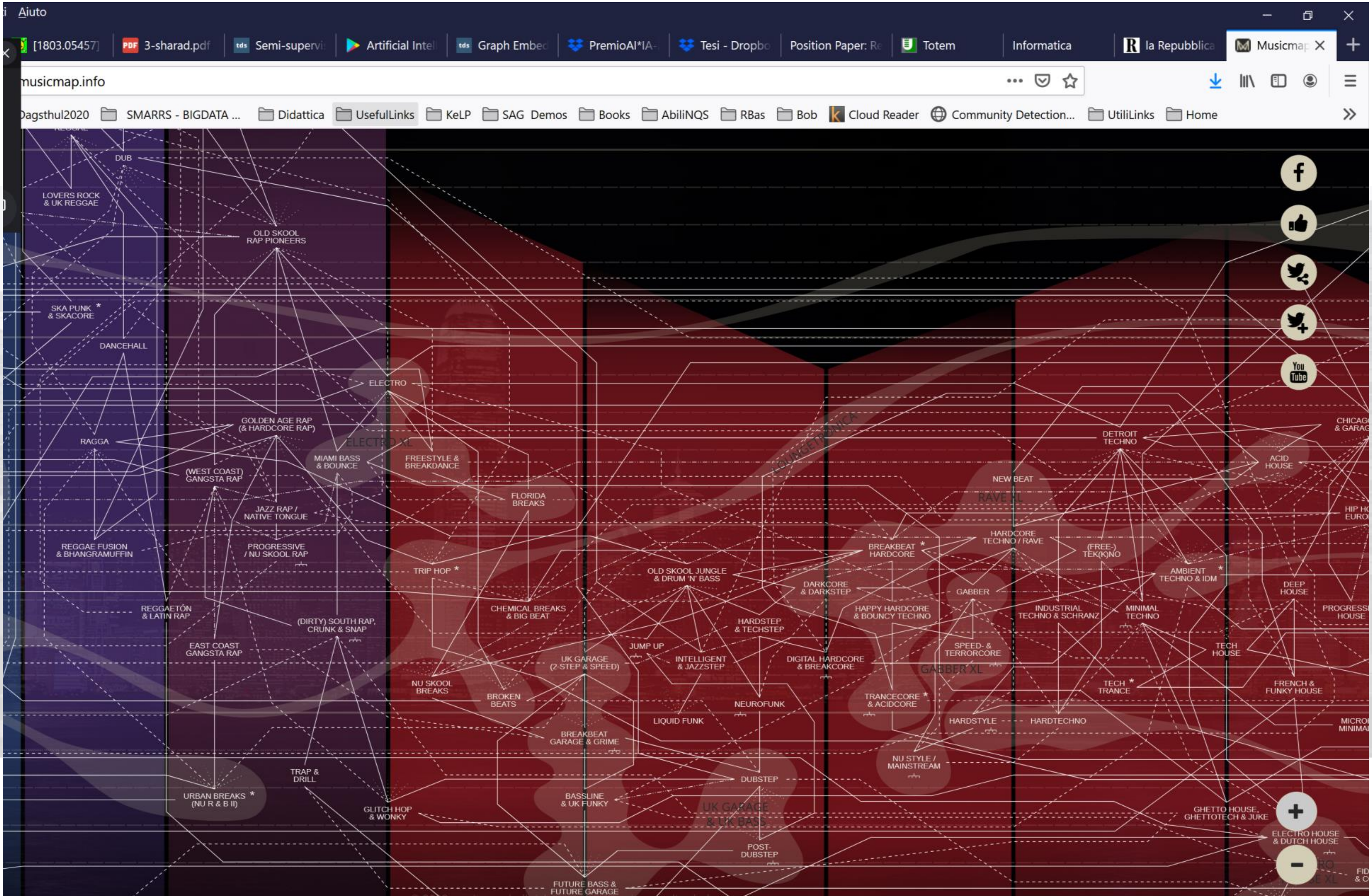
It has been discovered from data

No top-down design has been applied, as in knowledge engineering, but only bottom-up inferences (i.e. generalization from data)




nap
Front 242
Wisp
The Flashbulb
Coil
Alec Empire
Tangerine Dream
Atari Teenage Riot
Mouse On Mars
Biosphere
Venetian Snares
Dj Liquid
Dj Shadow
Speedy J
Binary Finary
Maurizio
Kraftwerk
Brian Eno
Monolake
The Orb
Alphawezen
Amon Tobin
Arovane
Stereolab
Shpongle
Future Sound Of London
My Bloody Valentine
Autechre
Squarepusher
Plaid
Orbital
Aphex Twin
Underworld
Boards Of Canada
Sonic Youth
Mum
Four Tet
Infected Mushroom
Skinny Puppy
Burial
e Vibert
Basement Jaxx
Chemical Brothers
Mogwai
The Prodigy
Björk
Animal Collective
Sigur Ros
The Mars Volta

Music maps: 2020



Information, Web and Natural Language

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

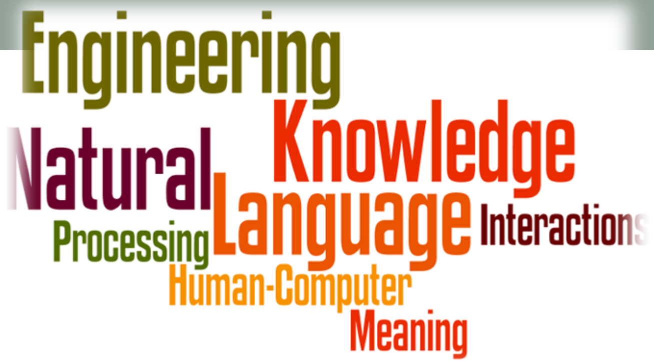
Information, Web and language

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

Who is Hu Jintao?

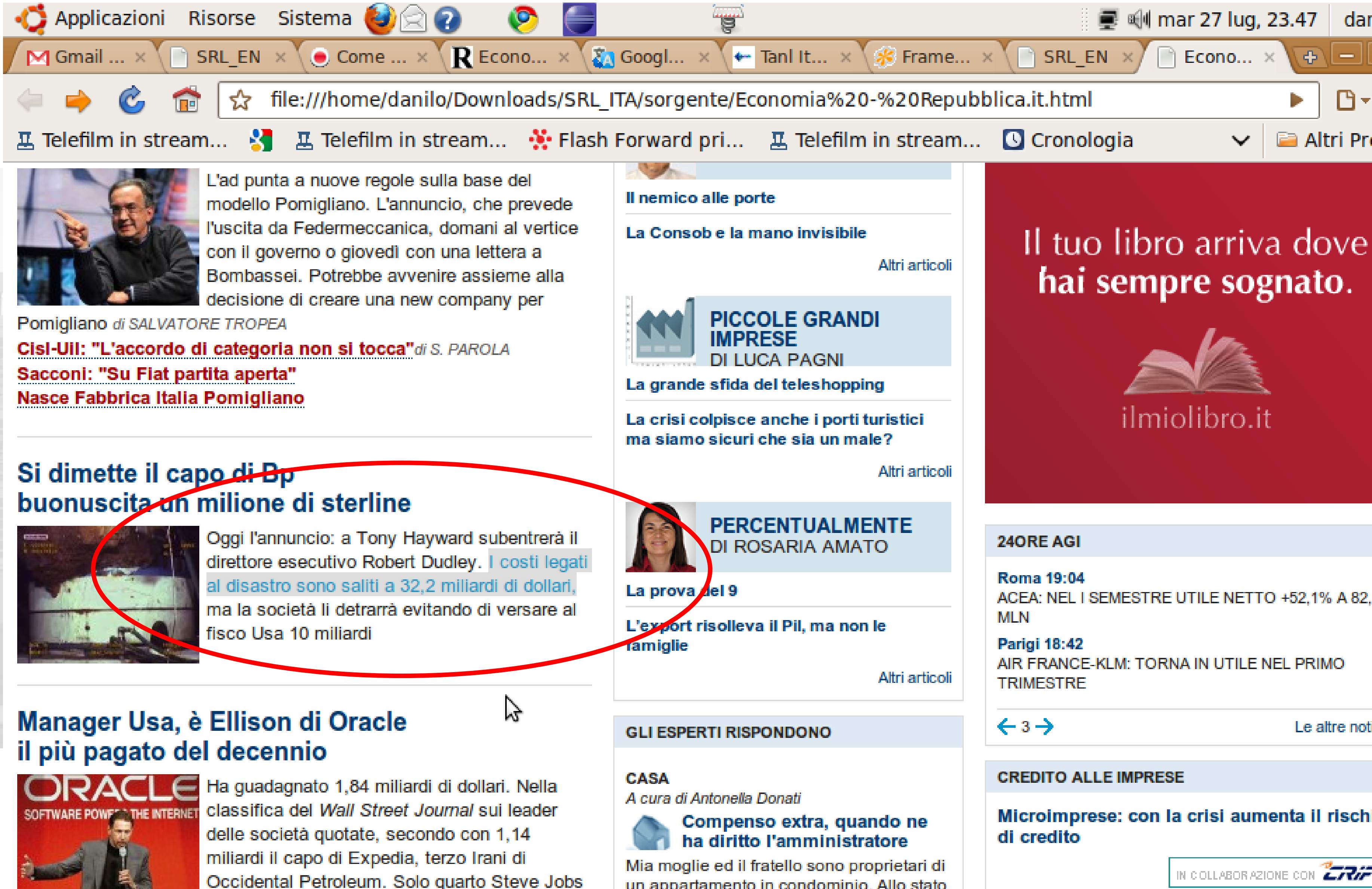

+Tu Web Immagini Video Mappe News Gmail Altro -
Hu Jintao
Ricerca
Circa 725.000 risultati (0,09 secondi)
Tutto
Mappe
Video
Notizie
Per argomento
Qualsiasi dimensione
Grandi
Medie
Icone
Maggiori di...
Dimensioni esatte...
Qualsiasi colore
A colori
Bianco e nero
Qualsiasi tipo
Volti
Foto
Clip art
Disegni
Visual. standard
Mostra dimensioni



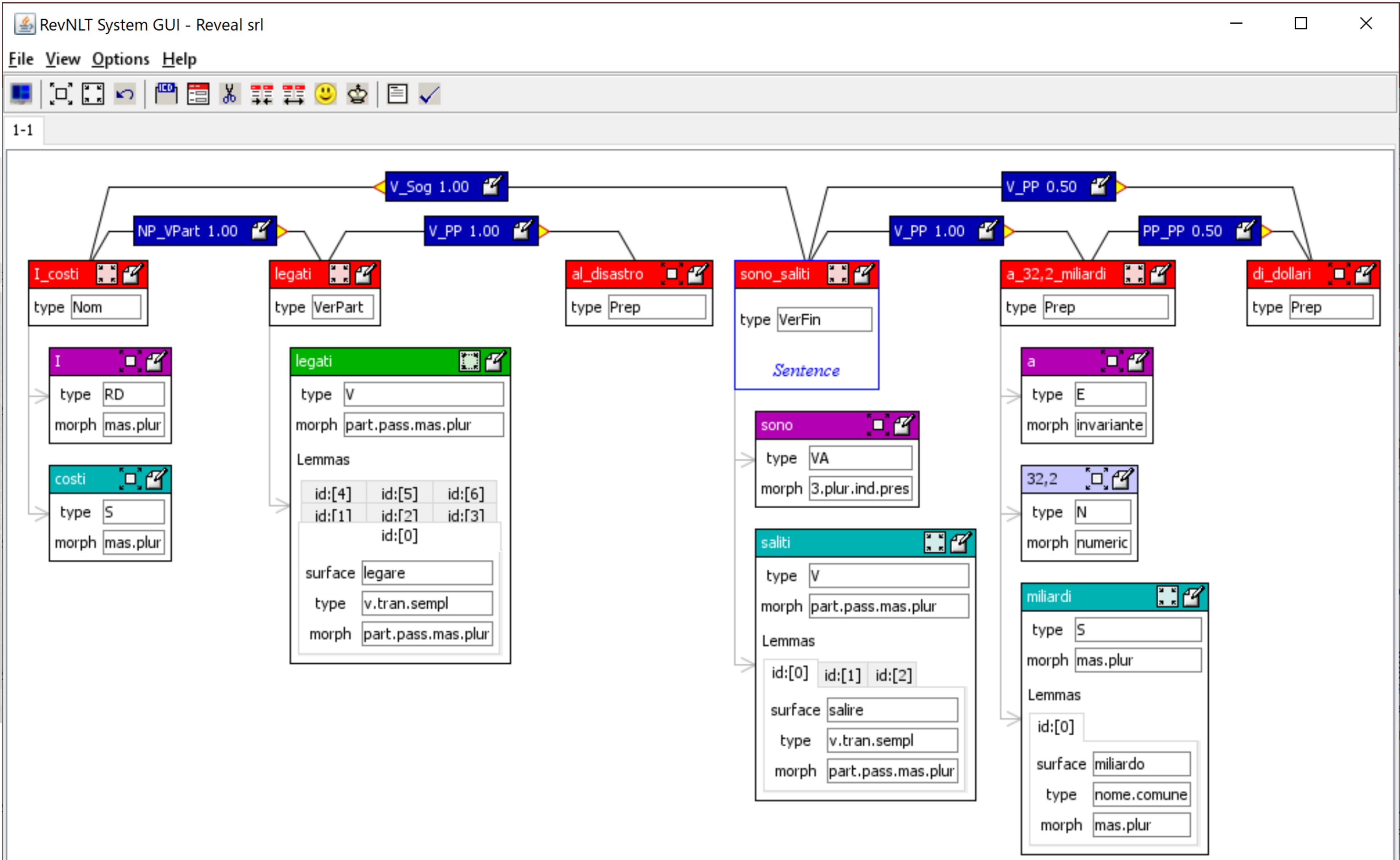
Content Semantics and Natural Language

- 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)

Semantics and News



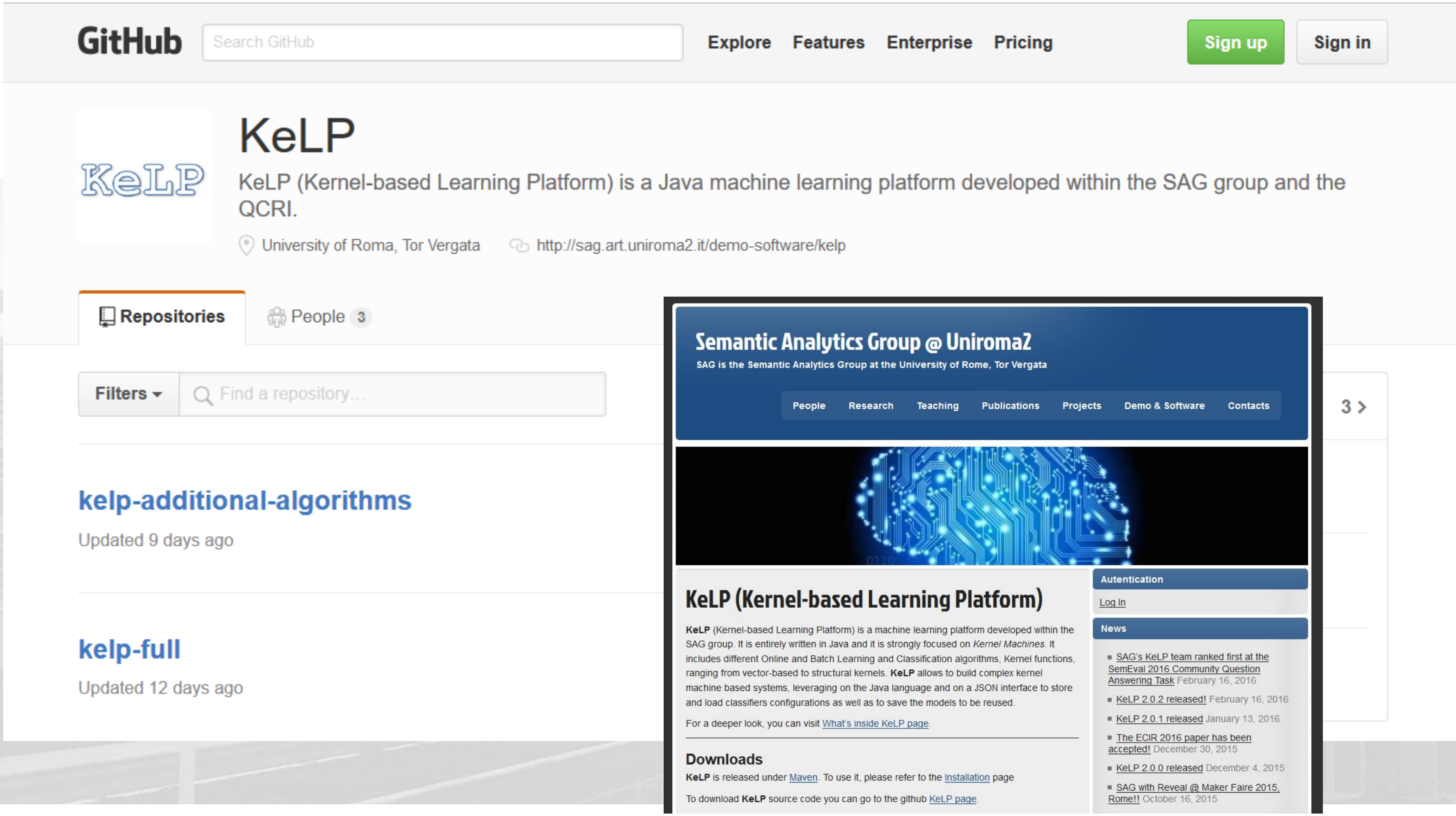
NLP: parsing Web texts



Course Laboratories:

- During the Course some laboratory sessions (4 hours per week) are scheduled jointly twith the *Machine Learning* course (Gambosi):
  - Machine Learning platforms: Pytorch, KELP, SciKit
  - NLP tools: Spacy, RevNLT (Reveal s.r.l.)
  - Deep Learning Tools for NLP:
    - Recursive Neural Networks for the acquisition of Domain specific Dictionaries
    - Transformers for Semantic Parsing and Natural Language Inference (e.g. paraphrasing)
    - Named Entity recognition and Wikification
  - Deep Learning for Visual Recognition
    - Object Detection, Captioning, Visual Question Answering
  - Deep Learning for Web Applicatuions:
    - Sentiment Analysis
    - Fake News Detection

Un esempio: Kelp: Java-based kernel framework



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"

General Description | Subtasks | Data and Tools

SemEval-2016 Task 3

Results

- The evaluation results can be found here
- The gold labels, submissions and scores for all teams can be foun
- The gold labels inside the test XML can be found here

Task participants are strongly encouraged to submit a system des
2016:
http://alt.qcri.org/semeval2016/index.php?id=call-for-papers

KELP applications: cQA

General Description | Subtasks | Data and Tools | Important Dates | Results | Call for Papers

SemEval-2016 Task 3

Results

- The evaluation results can be found here
- The gold labels, submissions and scores for all teams can be found here
- The gold labels inside the test XML can be found here

Task participants are strongly encouraged to submit a system description paper by March 4, 2016:
http://alt.qcri.org/semeval2016/index.php?id=call-for-papers

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*

Other Info

Announcements

KELP applications: cQA

Team ID	Team Affiliation
ConvKN	Qatar Computing Research Institute,
ECNU	East China Normal University, Chin
ICL00	Institute of Computational Lingustic
ICRC-HIT	Intelligence Computing Research Ce
ITNLP-AiKF	Intelligence Technology and Natural
Kelp	University of Roma, Tor Vergata, Ita
MTE-NN	Qatar Computing Research Institute,
overfitting	University of Waterloo, Canada
PMI-cool	Sofia University, Bulgaria
QAIIIT	IIIT Hyderabad, India
QU-IR	Qatar University, Qatar
RDI_team	RDI Egypt, Cairo University, Egypt
SemanticZ	Sofia University, Bulgaria
SLS	MIT Computer Science and Artificia
SUper_team	Sofia University, Bulgaria; Qatar Co
UH-PRHLT	Pattern Recognition and Human Lan
	Universitat Politècnica de València;
UniMelb	The University of Melbourne, Austr
UPC_USMBA	Universitat Politècnica de Catalunya

Table 5: The particip

	Submission	MAP	AvgRec	MRR	P	R	F1	Acc
1	**Kelp-primary**	**79.19_1**	**88.82_1**	**86.42_1**	**76.96_1**	**55.30_8**	**64.36_5**	**75.11_2**
	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
2	**ConvKN-primary**	**77.66_2**	**88.05_3**	**84.93_4**	**75.56_2**	**58.84_6**	**66.16_2**	**75.54_1**
3	**SemanticZ-primary**	**77.58_3**	**88.14_2**	**85.21_2**	**74.13_4**	**53.05_{10}**	**61.84_8**	**73.39_5**
	ConvKN-contrastive2	77.29	87.77	85.03	74.74	59.67	66.36	75.41
4	**ECNU-primary**	**77.28_4**	**87.52_5**	**84.09_6**	**70.46_6**	**63.36_4**	**66.72_1**	**74.31_4**
	SemanticZ-contrastive1	77.16	87.73	84.08	75.29	53.20	62.35	73.88
5	**SUper_team-primary**	**77.16_5**	**87.98_4**	**84.69_5**	**74.43_3**	**56.73_7**	**64.39_4**	**74.50_3**
	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
	SLS-contrastive1	76.46	87.47	83.27	60.09	69.68	64.53	68.87
6	**MTE-NN-primary**	**76.44_6**	**86.74_7**	**84.97_3**	**56.28_9**	**76.22_1**	**64.75_3**	**66.27_8**
7	**SLS-primary**	**76.33_7**	**87.30_6**	**82.99_7**	**60.36_8**	**67.72_3**	**63.83_6**	**68.81_7**
	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
	ICRC-HIT-contrastive1	73.34	84.81	79.65	63.43	69.30	66.24	71.28
8	**ITNLP-AiKF-primary**	**71.52_8**	**82.67_9**	**80.26_8**	**73.18_5**	**19.71_{12}**	**31.06_{12}**	**64.43_9**
	ECNU-contrastive1	71.34	83.39	78.62	66.95	41.31	51.09	67.86
9	**ICRC-HIT-primary**	**70.90_9**	**83.36_8**	**77.38_{10}**	**62.48_7**	**62.53_5**	**62.50_7**	**69.51_6**
10	**PMI-cool-primary**	**68.79_{10}**	**79.94_{10}**	**80.00_9**	**47.81_{12}**	**70.58_2**	**57.00_9**	**56.73_{12}**
	UH-PRHLT-contrastive1	67.57	79.50	77.08	54.10	50.11	52.03	62.45
11	**UH-PRHLT-primary**	**67.42_{11}**	**79.38_{11}**	**76.97_{11}**	**55.64_{10}**	**46.80_{11}**	**50.84_{11}**	**63.21_{10}**
	UH-PRHLT-contrastive2	67.33	79.34	76.73	54.97	49.13	51.89	62.97
12	**QAIIIT-primary**	**62.24_{12}**	**75.41_{12}**	**70.58_{12}**	**50.28_{11}**	**53.50_9**	**51.84_{10}**	**59.60_{11}**
	QAIIIT-contrastive2	61.93	75.22	69.95	49.48	49.96	49.72	58.93
	QAIIIT-contrastive1	61.80	75.12	69.76	49.85	50.94	50.39	59.24
	Baseline 1 (IR)	**59.53**	**72.60**	**67.83**	—	—	—	—
	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	**57.80**	40.64
	Baseline 4 (all 'false')	—	—	—	—	—	—	**59.36**

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 primary runs with respect to the evaluation measure in the respective column.

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