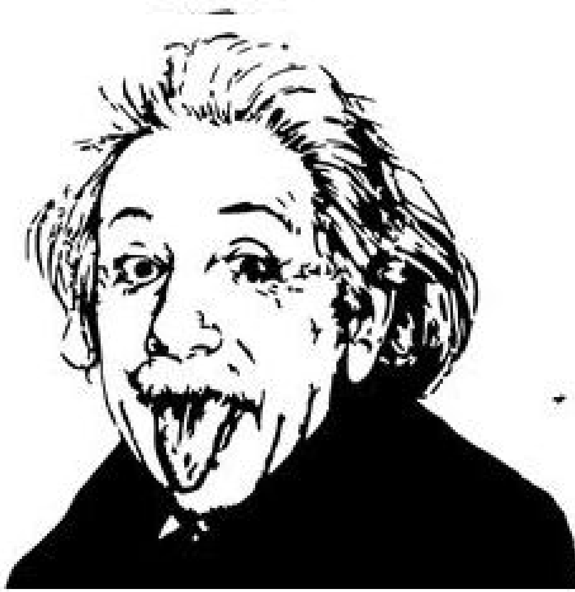
V.I.TECH

Software research and development

Be on Science's Pulse with NLP
or
"What are the serious men talking about?"



Kodliuk Tetiana, Data Scientist

Who are you?
Why did we wake up so early?

V.I.TECH

Data Scientist

Active Wizards

Lead of Data Scientists

Aegis
SCHOOL OF BUSINESS
SCHOOL OF TELECOMMUNICATION

Lecturer in Apache Spark

ТНПУ
ТЕРНОПІЛЬСЬКИЙ НАЦІОНАЛЬНИЙ ПЕДАГОГІЧНИЙ УНІВЕРСИТЕТ

НАЦІОНАЛЬНА
АКАДЕМІЯ
НАУК
УКРАЇНИ
1918

Lecturer in Math
Ph. D. in Math

Analyst

We are living in magnificent time!!!

According to futuretimeline.net:

2050
- Robots take 50% of our jobs

2100
- Human intelligence is being vastly amplified by AI

2150
- Terraforming of Mars is underway

2200
- Traditional employment is becoming obsolete

4000
- Computer science is reaching its ultimate potential

Artificial Intelligence is everywhere, isn't it?

image recognition

knowledge management

probabilistic reasoning

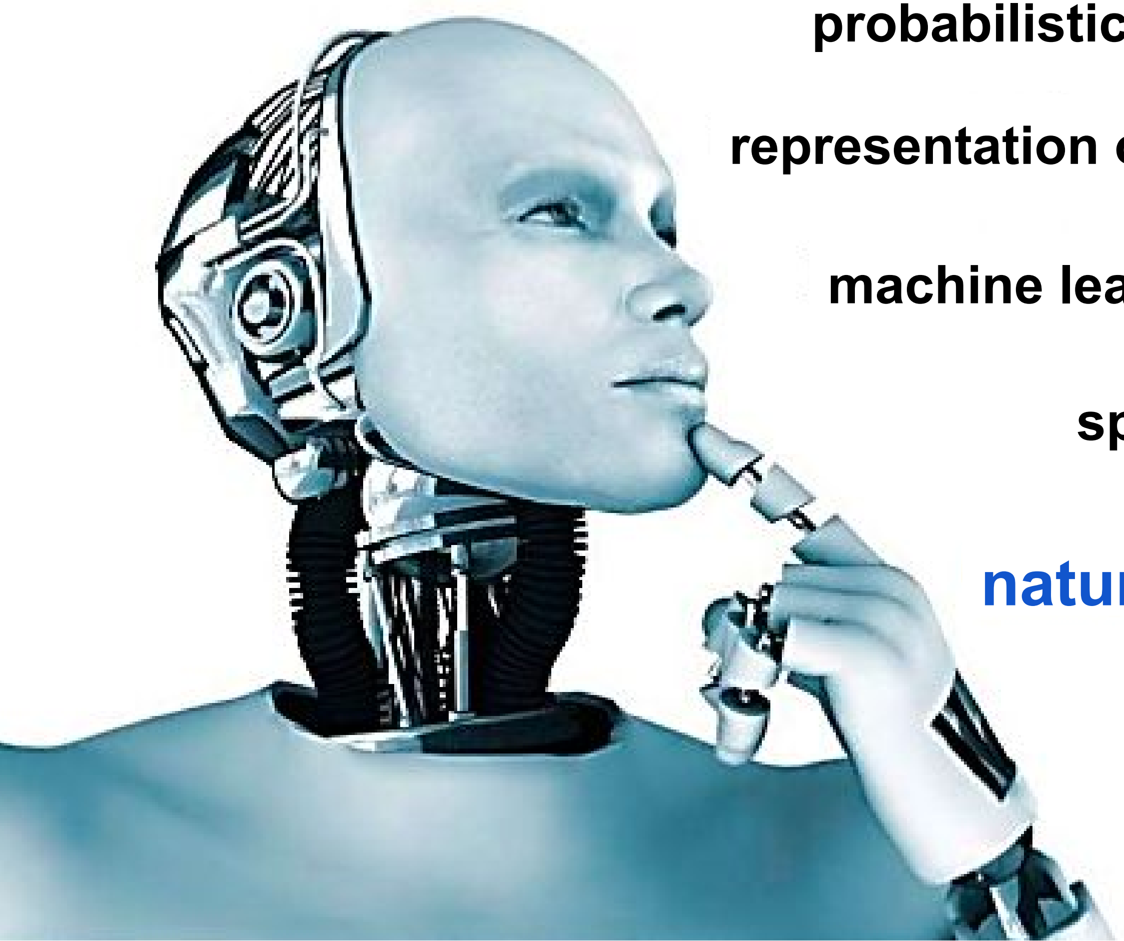
representation of human expression

machine learning

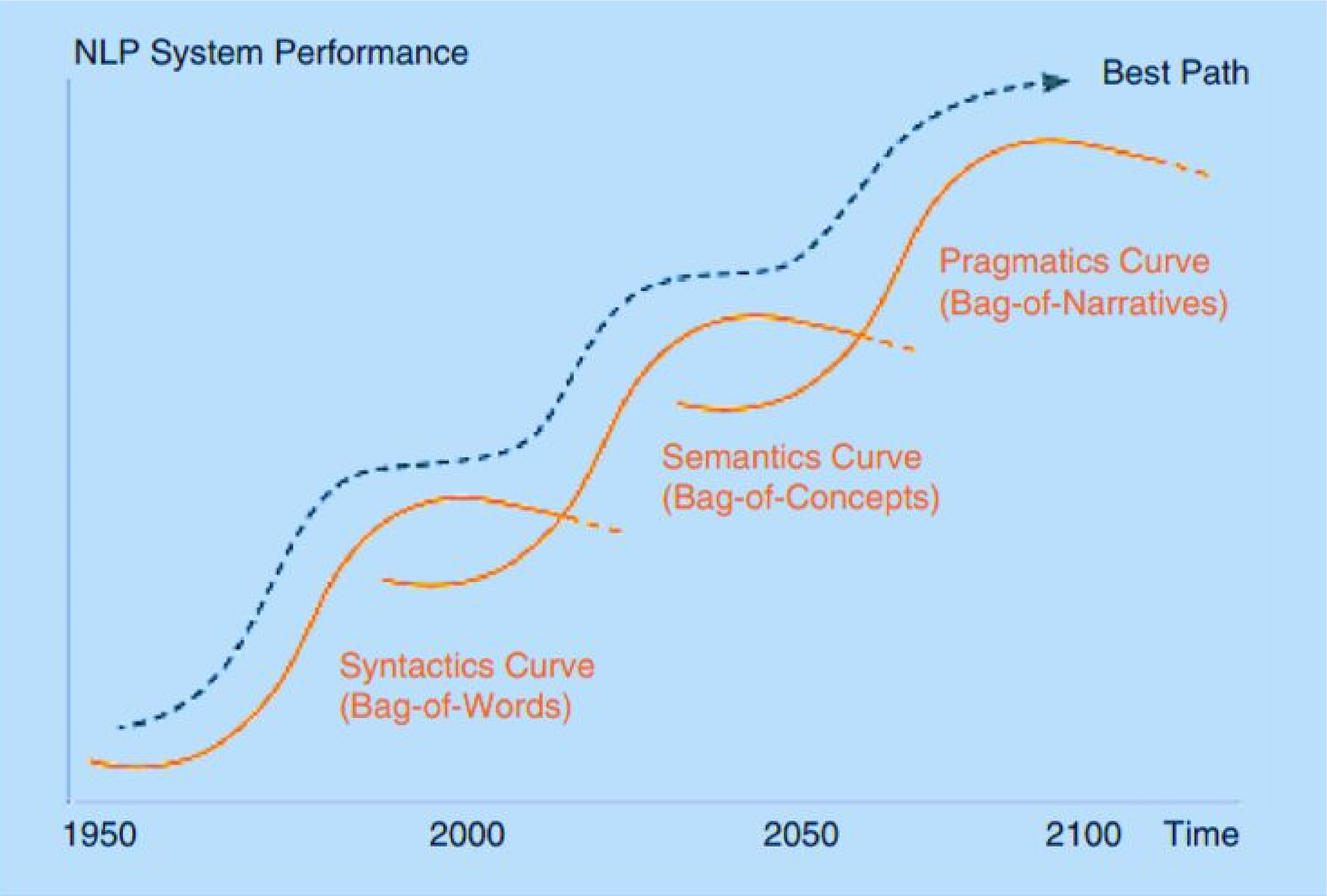
speech to text

natural language processing

robotics



Jumping NLP Curves (Stanford, 2014)



http://sentic.net/jumping-nlp-curves.pdf

V.I.TECH

So our idea was born...

Extract the trends
from the scientific publications

https://www.ucl.ac.uk/human-evolution/

Keywords extraction

Keyword (keyphrase) extraction
is tasked with the automatic identification of terms (phrases) that best describe the subject of a document

www.socialappshq.com

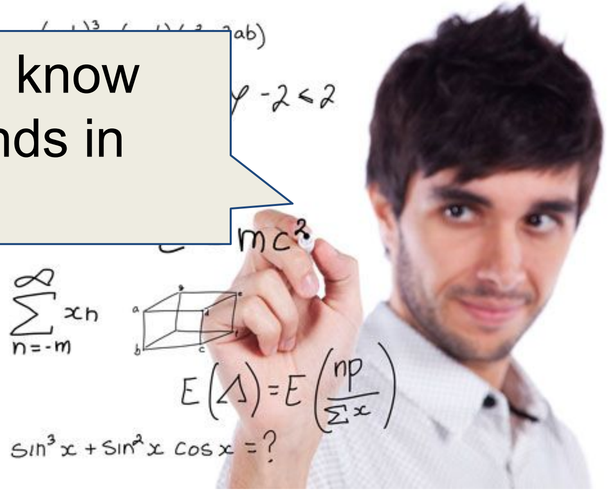
Everyone needs it...

I want to know only KEY news.

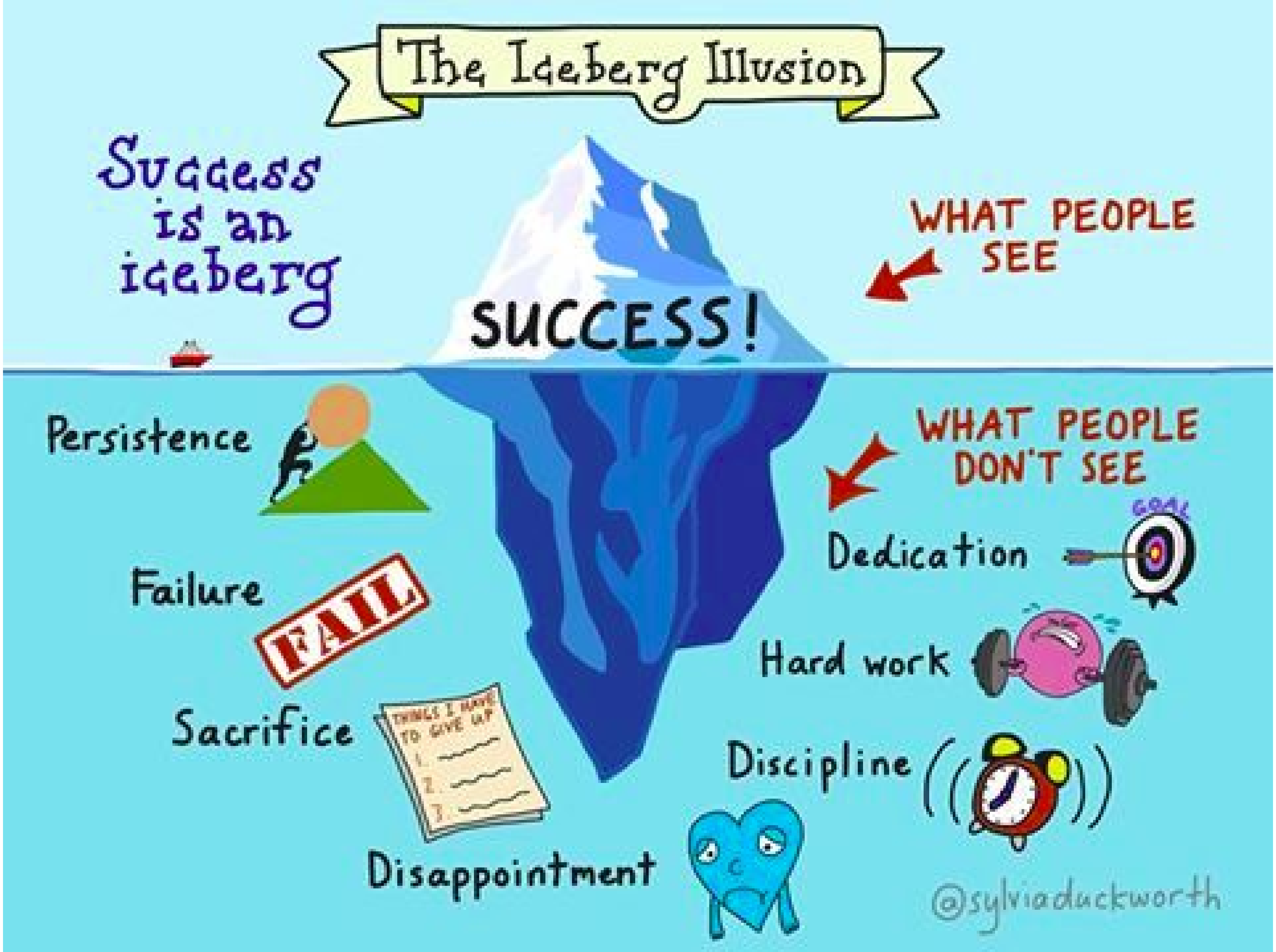
I want to know KEY approaches in medicine.

I want to now KEY problems of customers.

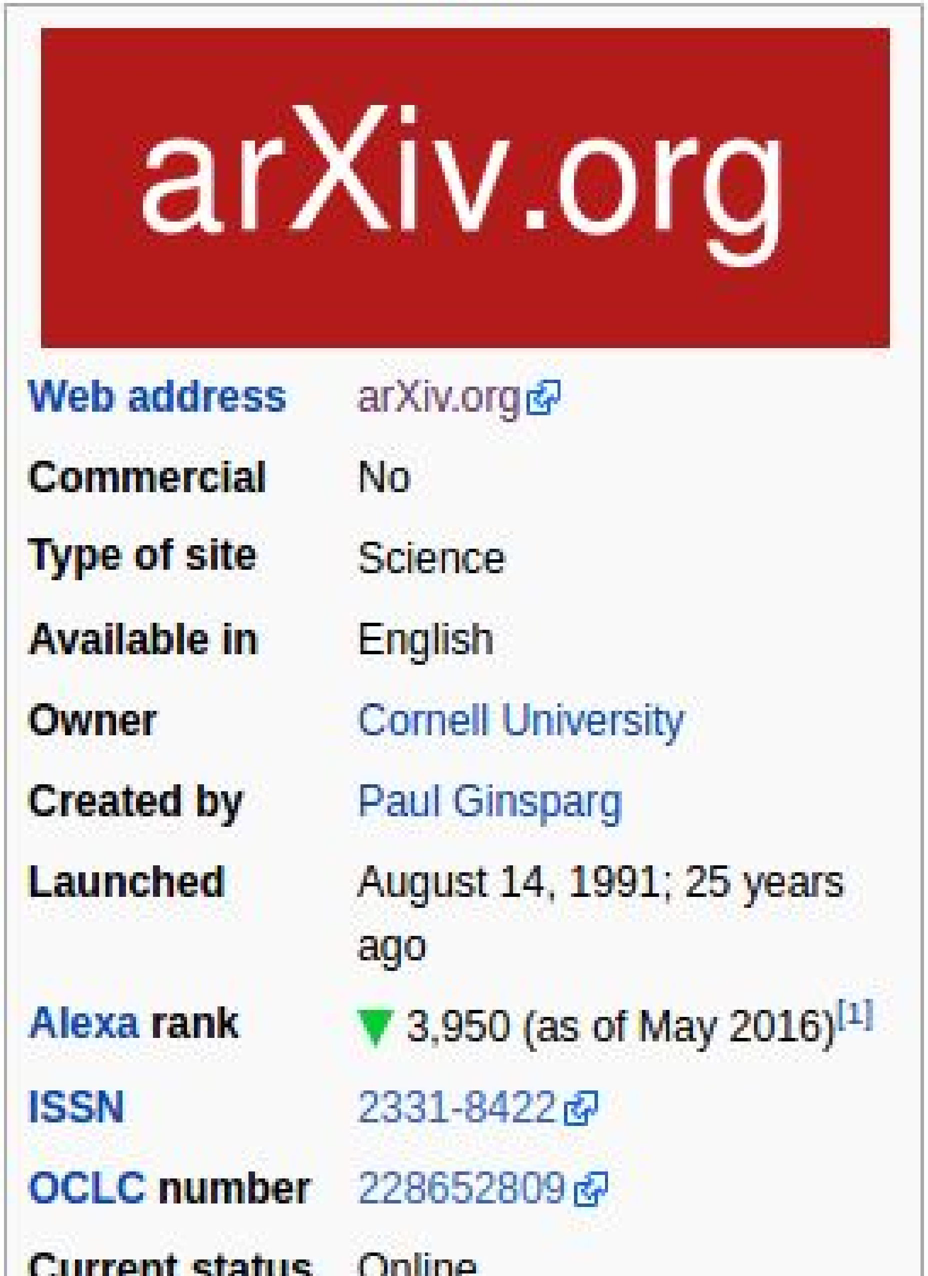
I want to know KEY trends in Science.


ab)
-2<2
mc²
$\sum_{n=-m}^{\infty} x_n$
$E(\Delta)=E\left(\frac{np}{\Sigma x}\right)$
$\sin^3 x + \sin^2 x \cos x = ?$

Let’s talk about a rough way



Corpus



arXiv.org

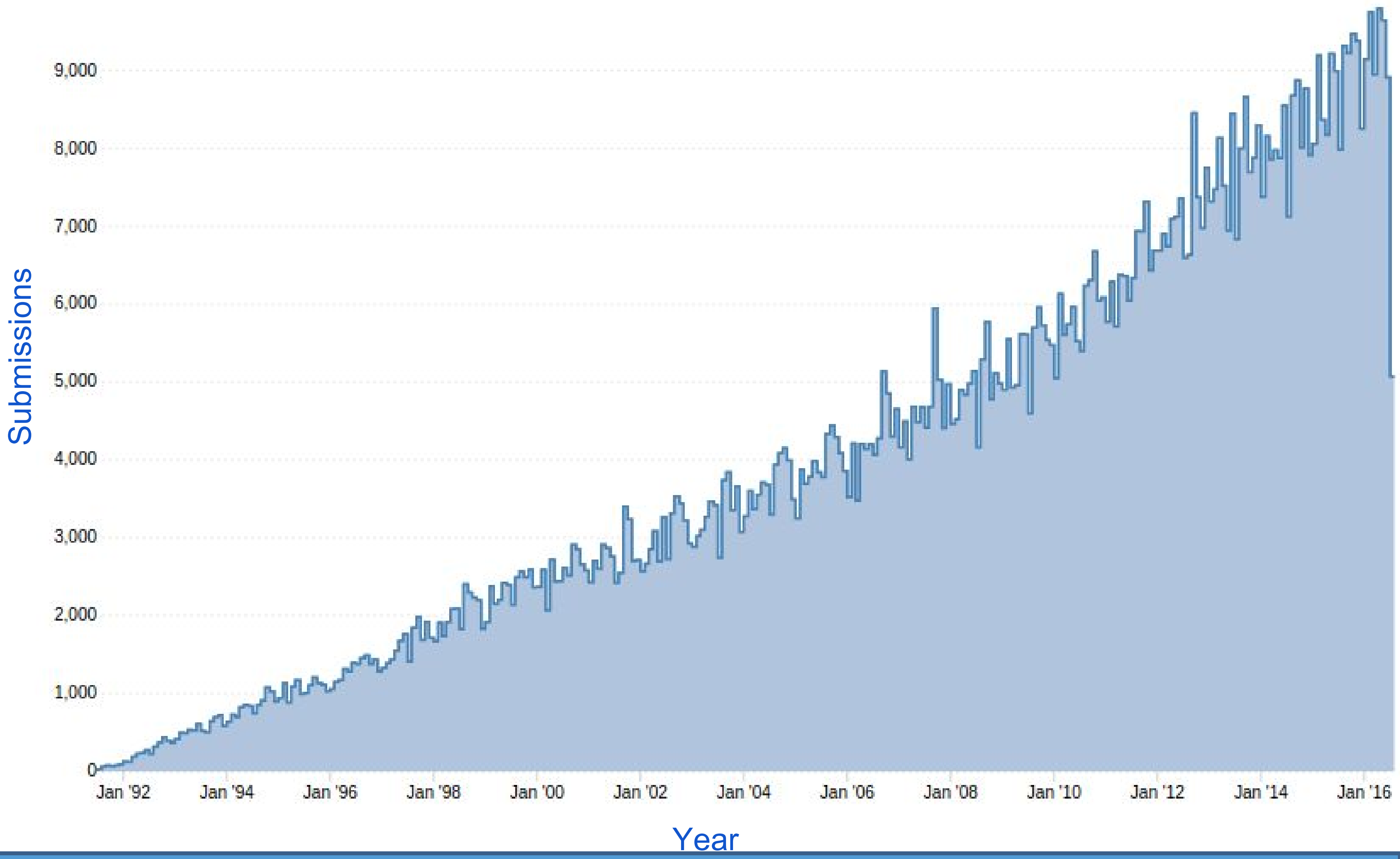
Web address	arXiv.org
Commercial	No
Type of site	Science
Available in	English
Owner	Cornell University
Created by	Paul Ginsparg
Launched	August 14, 1991; 25 years ago
Alexa rank	▼ 3,950 (as of May 2016)[1]
ISSN	2331-8422
OCLC number	228652809
Current status	Online

wikipedia.org

ArXiv Categories

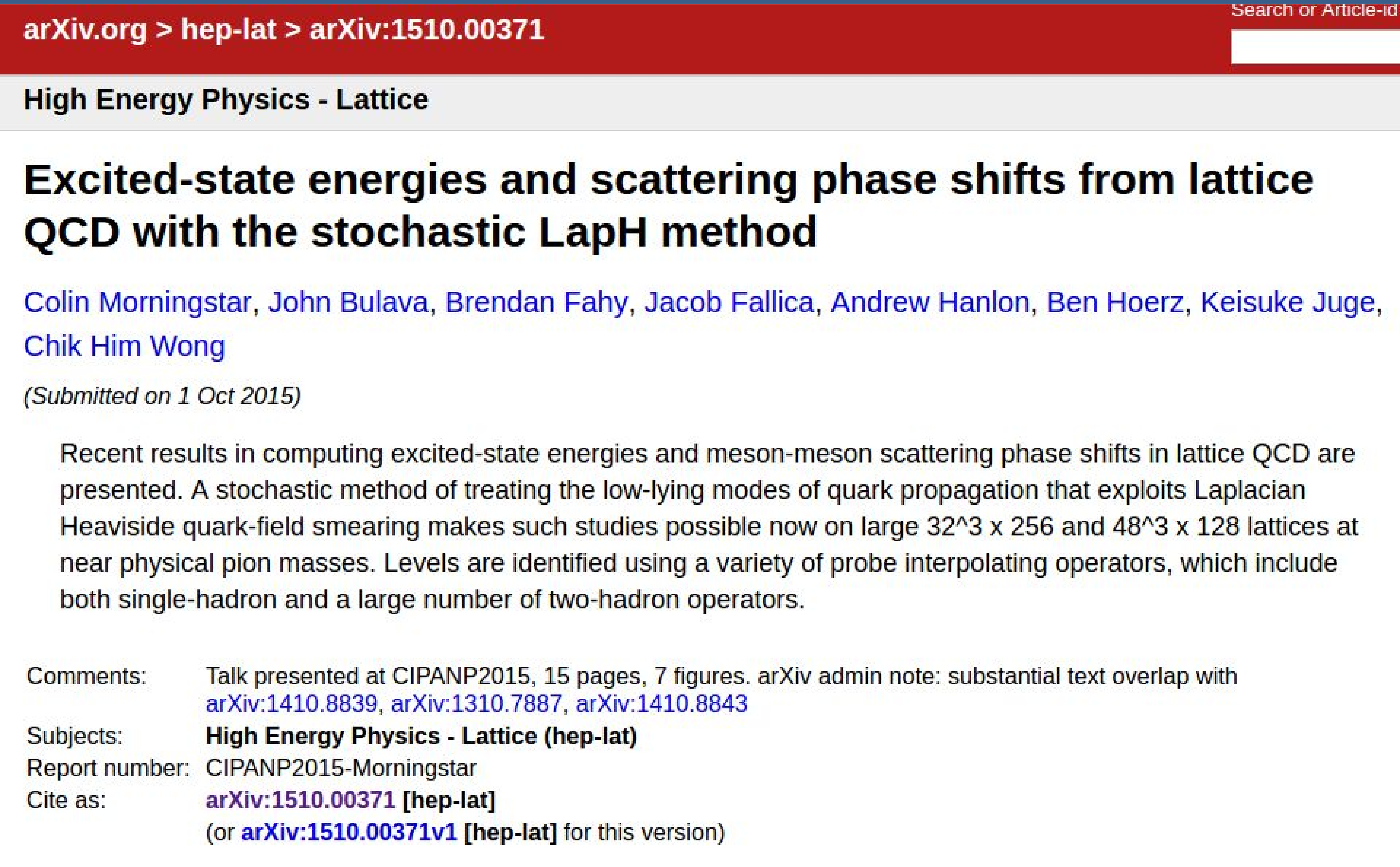
	Category	Number of subcategories
1	Statistics	5
2	Quantitative Biology	10
3	Computer Science	36
4	Nonlinear Sciences	5
5	Mathematics	32
6	Physics	39

ArXiv monthly submissions



http://arxiv.org/stats

How does the input data look like?



arXiv.org > hep-lat > arXiv:1510.00371

High Energy Physics - Lattice

Excited-state energies and scattering phase shifts from lattice QCD with the stochastic LapH method

Colin Morningstar, John Bulava, Brendan Fahy, Jacob Fallica, Andrew Hanlon, Ben Hoerz, Keisuke Juge, Chik Him Wong

(Submitted on 1 Oct 2015)

Recent results in computing excited-state energies and meson-meson scattering phase shifts in lattice QCD are presented. A stochastic method of treating the low-lying modes of quark propagation that exploits Laplacian Heaviside quark-field smearing makes such studies possible now on large 32^3 x 256 and 48^3 x 128 lattices at near physical pion masses. Levels are identified using a variety of probe interpolating operators, which include both single-hadron and a large number of two-hadron operators.

Comments:	Talk presented at CIPANP2015, 15 pages, 7 figures. arXiv admin note: substantial text overlap with arXiv:1410.8839, arXiv:1310.7887, arXiv:1410.8843
Subjects:	**High Energy Physics - Lattice (hep-lat)**
Report number:	CIPANP2015-Morningstar
Cite as:	**arXiv:1510.00371 [hep-lat]** (or **arXiv:1510.00371v1 [hep-lat]** for this version)

Where you can find the Data?

Bulk Metadata Access

- ArXiv API
- OAI-PMH
- RSS

It was pretty easy, but...

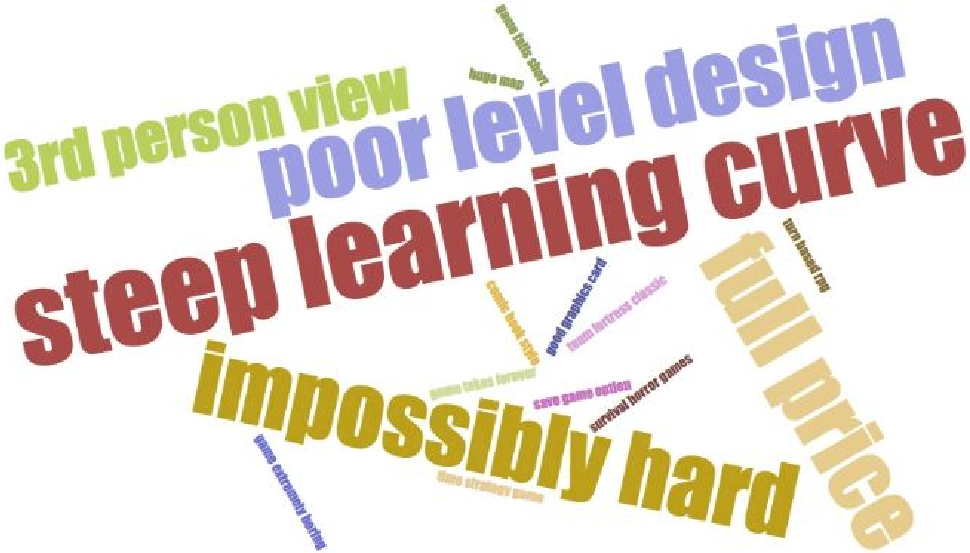
Problems

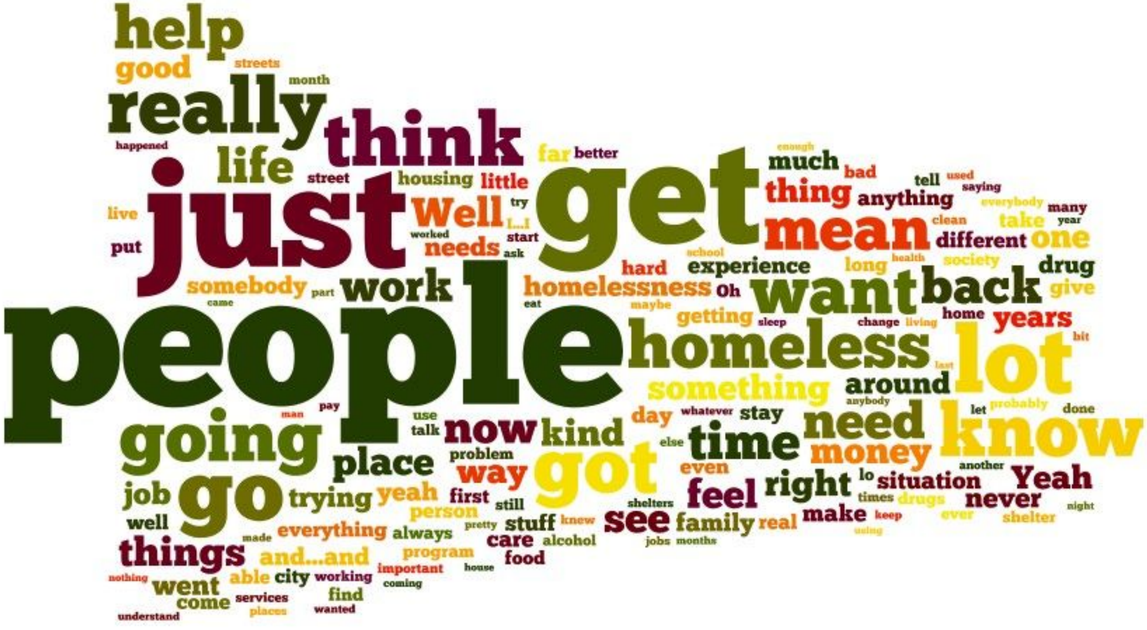
- ➔ The versions of the papers can change at any moment
- ➔ Arxiv.org doesn’t allow to scrape a lot of atoms at the same time
- ➔ It can be the problem with an internet
- ➔ It can be the extra characters in abstracts, author names, title etc.
- ➔ The categories names can change: "adap-org" = "nlin.AO“, "Q-alg" = "math.QA“
- ➔ Ids of papers can have different format: 1606.04426, 160323

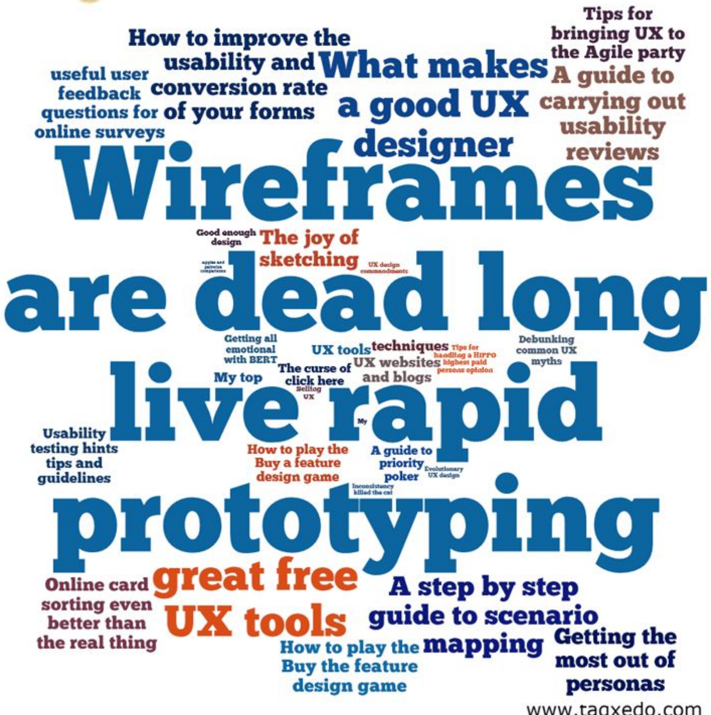
Solutions

- ★ We update the versions of papers each month
- ★ We are waiting 20 second for scraping new atoms
- ★ We are retrying after 30 seconds each time.
- ★ We use encoding
- ★ We learn the changes and map them

Keywords or keyphrases?

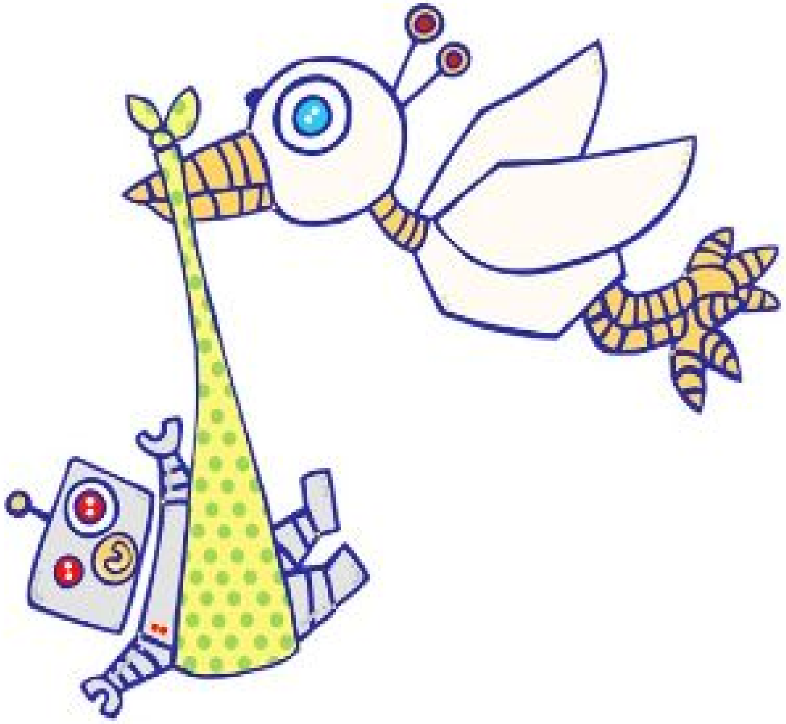


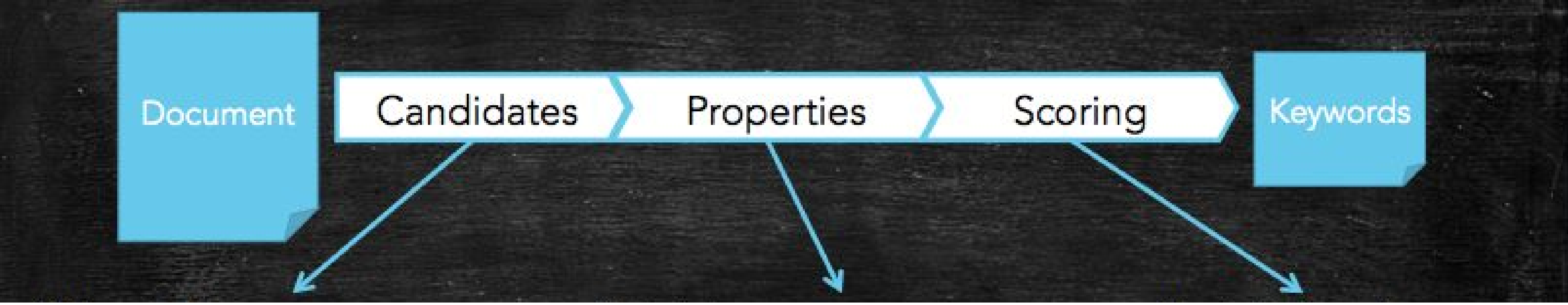




Actual methods for keyphrases extraction

- Statistical methods: Frequency, TF-IDF, BM25
- RAKE
- TextRank, KeyRank
- Supervised Machine Learning
- Neural Networks





Words
N-grams, phrases
POS patterns
Named entities
….

Frequencies
Weights: TF-IDF, BM25
Rank
….

Manual
Supervised
Depended
….

Rapid Automatic Keyword Extraction (RAKE)

alamy

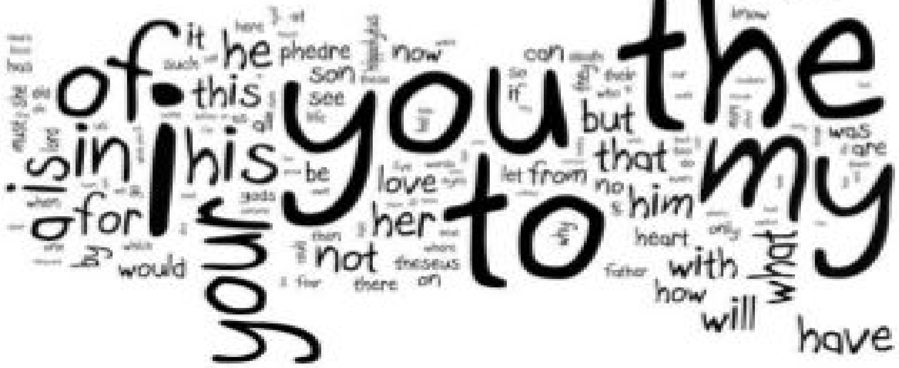
Author: Stuart Rose (2010)

- Unsupervised method for extracting keywords
- Incorporate cooccurrence and frequency of words

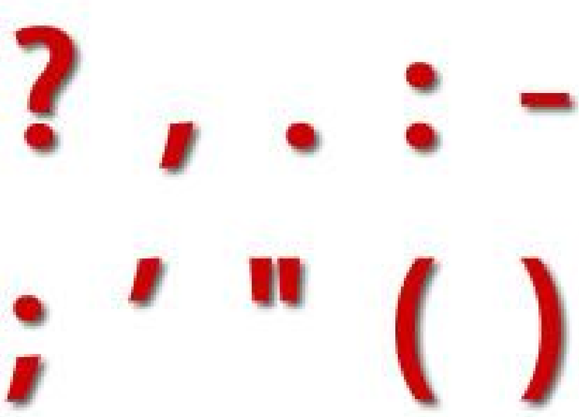
Candidate keywords selecting

RAKE partitions the text by using

stop words



phrase delimiters



Candidate keywords selecting

Compatibility of systems of linear constraints over the set of natural numbers. Criteria of compatibility of a system of linear Diophantine equations, strict inequations, and nonstrict inequations are considered. Upper bounds for components of a minimal set of solutions and algorithms of construction of minimal generating sets of solutions for all types of systems are given. These criteria and the corresponding algorithms for constructing a minimal supporting set of solutions can be used in solving all the considered types of systems and systems of mixed types.

Compatibility – systems – linear constraints – set – natural numbers – Criteria – compatibility – system – linear Diophantine equations – strict inequations – nonstrict inequations – Upper bounds – components – minimal set – solutions – algorithms – minimal generating sets – solutions – systems – criteria – corresponding algorithms – constructing – minimal supporting set – solving – systems – systems

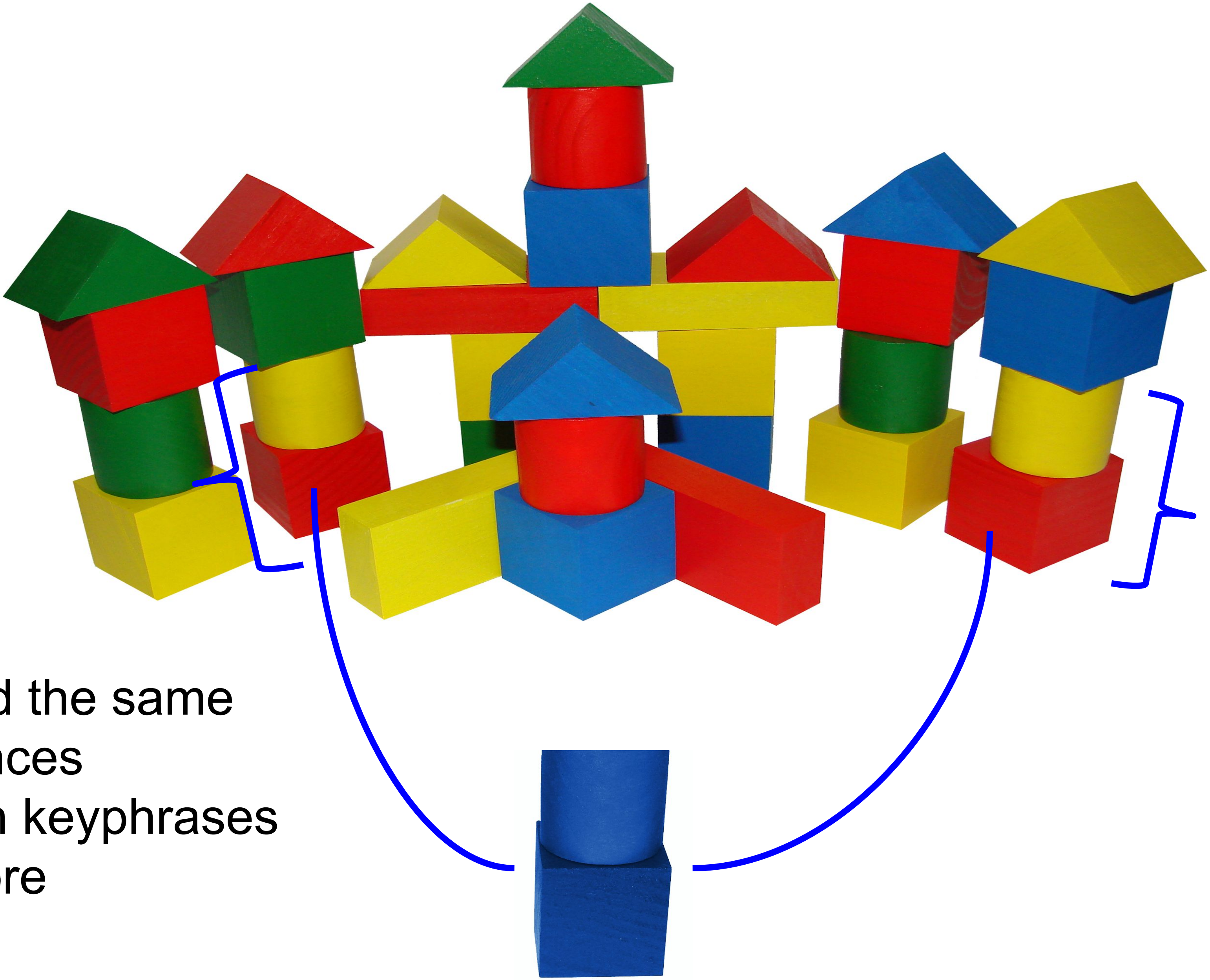
Candidates scoring

Score of phrase = SUM(score(word))

Metrics for calculating word scores:

1. word frequency: freq(w),
2. word degree: deg(w),
3. ratio of degree to frequency: deg(w)/freq(w)

Additional option: adjoining keywords



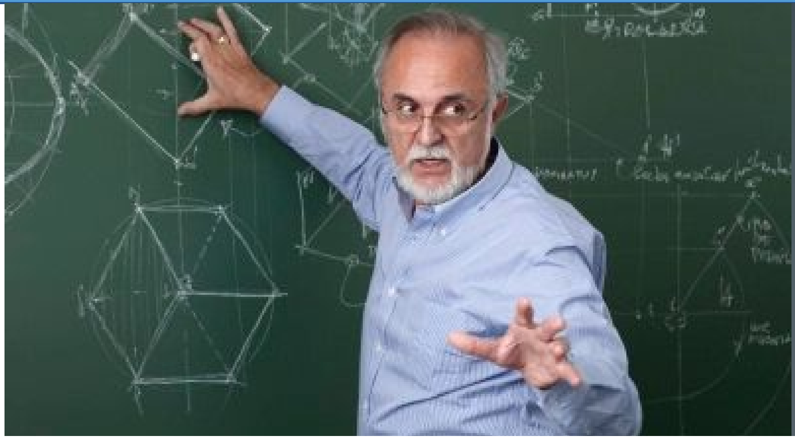
1. Find the same sequences
2. Join keyphrases
3. Score

Where am I?

I thought, she will speak about serious men...

Some strange tables and no word about men...

RAKE



Keyphrase	RAKE
spectral amplitude coding optical code division multiple access networks intelligent pinning based cooperative secondary control	51.64
service function localization enabling fine grained rdf data completeness assessment balanced ranking mechanisms convolutional neural networks	51.33
strongly magnetized neutron stars powering superluminous supernovae remarkable magnetostructural coupling	49.85
randomized version room temperature tetragonal noncollinear antiferromagnet ptmnga optimal system maneuver	49.62

Keyphrase	RAKE
мінімум волі за мінімум долі	10.56
вона хотіла зніматись в кіно	10.04
він мені свої пісні співав	8.83
я-а-а почую голос твій	7.72
мало-мало-мало мені	7.72
вісім днів він її шукав	7.07
усі знайомі ледь знайомі	7.07
ключ не підійшов а може й не посмів	7.07
я слухав звук дощу і Біллі Холідей	7.07

"The best way to have a good idea is to have **a lot of ideas.**"

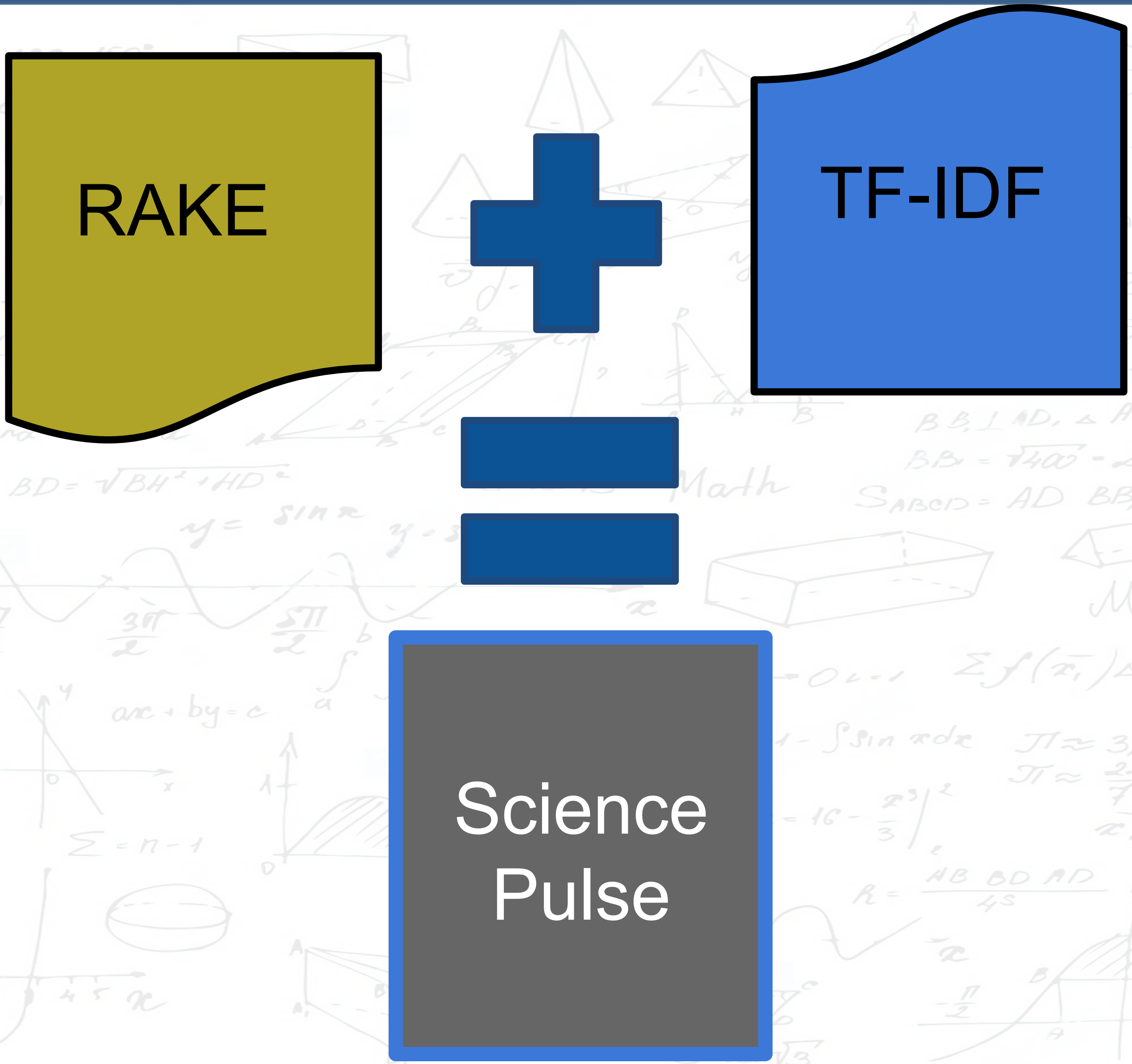
~ Linus Pauling

PersonalExcellence.co

What is the best method?

Method	Advantage	Disadvantage
TF-IDF	Important keyphrases extraction, n-grams possible	Candidates extraction
TextRank	Cooccurrences calculating	Long phrases
RAKE	Frequences, Coocurences calculating, candidates extraction	Long phrases
Supervised ML	High score of extraction	Train dataset is needed

Our Approach

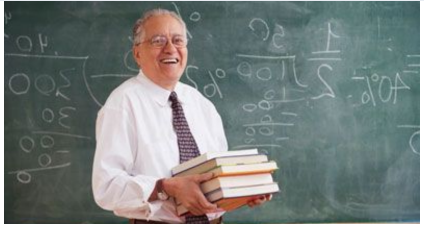


RAKE

TF-IDF

Science Pulse

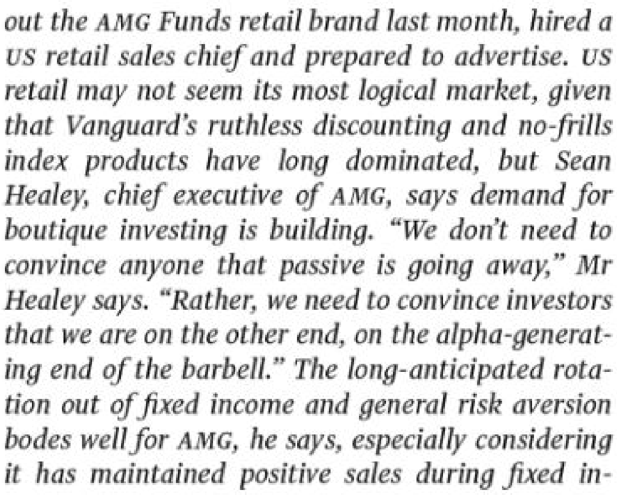
RAKE+TF-IDF



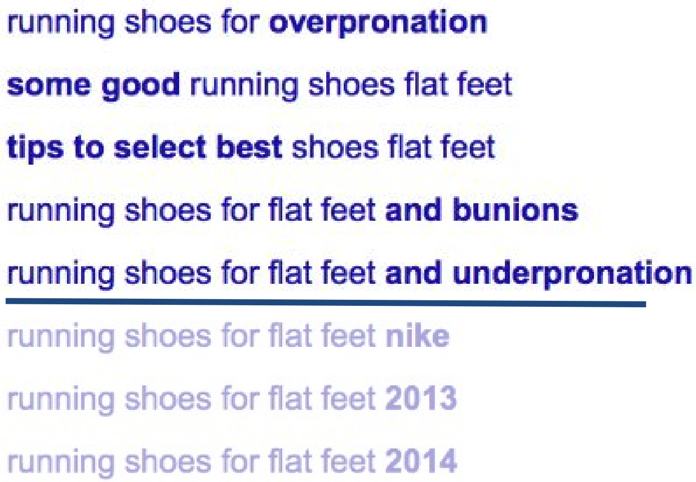
Keyphrase	Weight	Keyphrase	Weight
massive multiple input output	7,25	все буде добре	1
long short term memory architecture	5,12	коли тебе нема	1
Live action virtual reality games	3,15	небо над дніпром	1
low rank hankel matrix completion	3,04	хочу напитись тобою	0,78
multi point wireless energy transmission	3,01	жити без мети	0,78
tree augmented naive bayes classifier	2,89	мила моя сьюзі	0,78
long short term memorized fusion	2,15	тінь твого тіла	0,75
fine grained entity type classification	1,51	коли настане день	0,75
high speed railway communication systems	1,27	кожну хвилину життя	0,75
partially observable markov decision process	1,13	коли тобі важко	0,75

How to join different methods?

Input text	RAKE	TF-IDF



out the AMG Funds retail brand last month, hired a US retail sales chief and prepared to advertise. US retail may not seem its most logical market, given that Vanguard's ruthless discounting and no-frills index products have long dominated, but Sean Healey, chief executive of AMG, says demand for boutique investing is building. “We don't need to convince anyone that passive is going away,” Mr Healey says. “Rather, we need to convince investors that we are on the other end, on the alpha-generating end of the barbell.” The long-anticipated rotation out of fixed income and general risk aversion bodes well for AMG, he says, especially considering it has maintained positive sales during fixed in-



running shoes for **overpronation**

some good running shoes flat feet

tips to select best shoes flat feet

running shoes for flat feet **and bunions**

running shoes for flat feet **and underpronation**

running shoes for flat feet **nike**

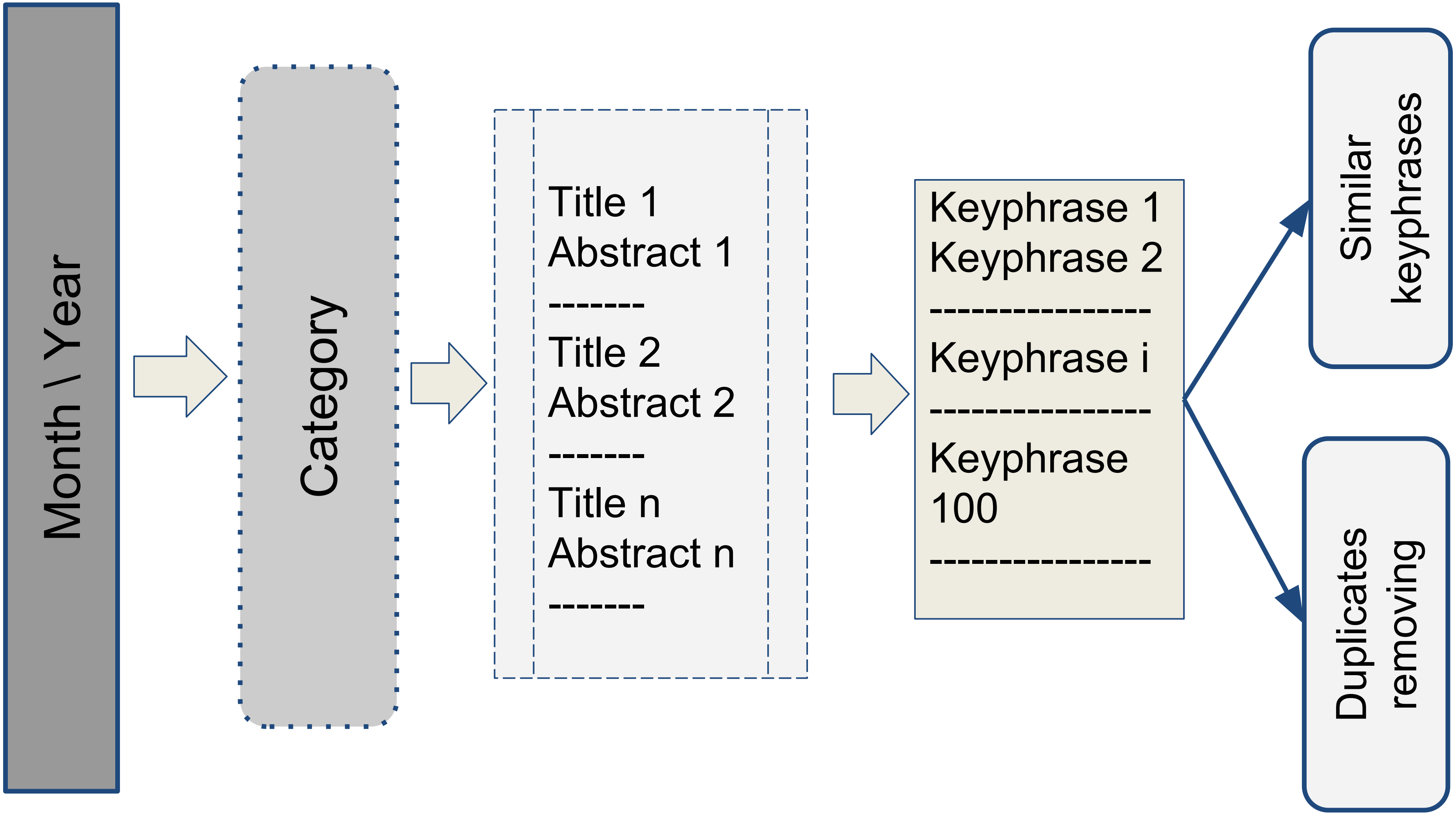
running shoes for flat feet **2013**

running shoes for flat feet **2014**

Keyphrase 1
Keyphrase 2
Keyphrase 3
Keyphrase 4
Keyphrase 5

	RAKE weight	TF-IDF score

Text for keyphrases selecting



Month \ Year

Category

Title 1
Abstract 1

Title 2
Abstract 2

Title n
Abstract n

Keyphrase 1
Keyphrase 2

Keyphrase i

Keyphrase 100

Similar keyphrases

Duplicates removing

Duplicates removing

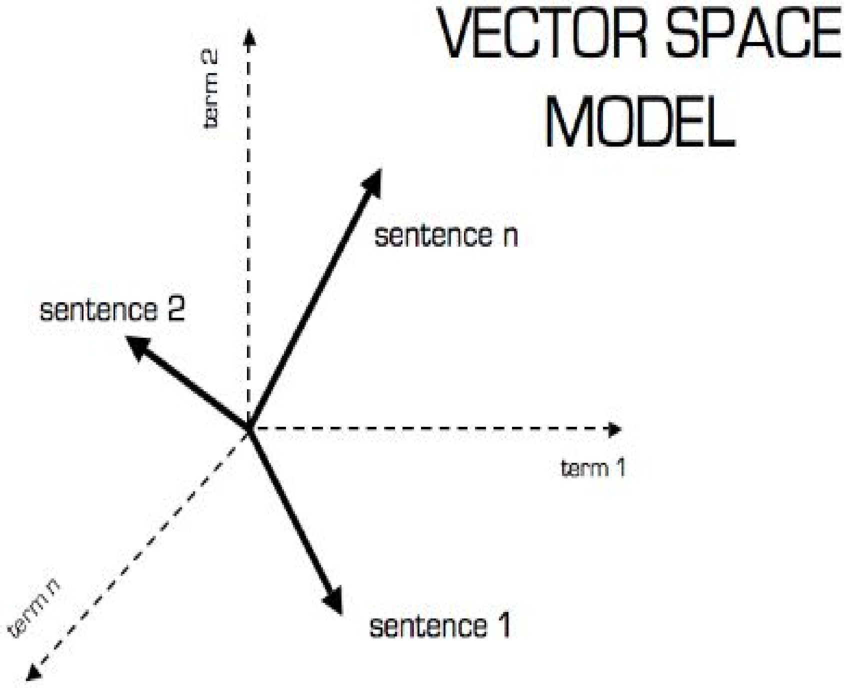
Bag of Words

Keyphrase 1
Keyphrase 2

Keyphrase n

Keyphrase 100

	Occurence in Keyphrase 1	Occurence in Keyphrase 2
word1	1	1
word2	0	0
wordn	1	0



cosine_similarity = cosine_sim(keyphrase1, keyphrase2)

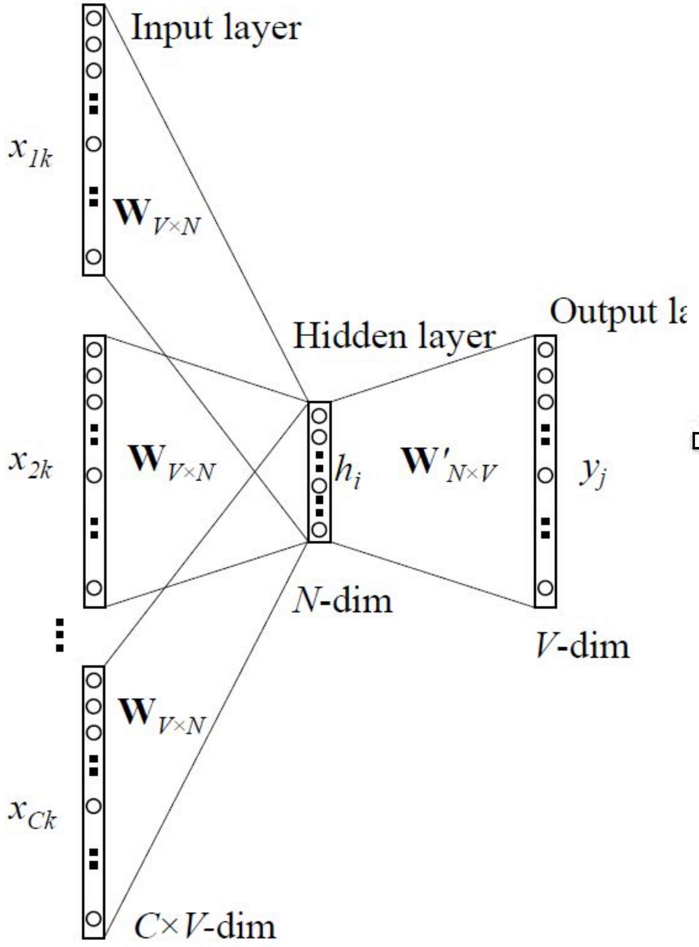
filter(similarity >= 0.8)

Trends Recommendations

Word2Vec

word

The
Cardinals
will
win
the
world
series



vector

(0.12, 0.23, 0.56)
(0.24, 0.65, 0.72)
(0.38, 0.42, 0.12)
(0.57, 0.01, 0.02)
(0.53, 0.68, 0.91)
(0.11, 0.27, 0.45)
(0.01, 0.05, 0.62)

stats.stackexchange.com

V.I.TECH

Recommendations: Word2Vec

Wikipedia+Gigaword 5
Number of dimensions : 300
Windows size: 10
✓

Wikipedia
Number of dimensions : 1000
Windows size: 10
🚫

ArXiv (abstracts)
Number of dimensions : 300
Windows size: 10
🚫

Additional rules for similarity

1. The year was selected as period of similar statements searching
2. The cosine distance between sets of words is calculated.
3. The lowest cosine similarity between statements should be equal 0.70

Welcome to Science Pulse

Science Pulse analytics

Keyphrase-Atom relationships

Mean number of Atoms per Keyphrase = 1.6178478064

Max number of Atoms per Keyphrase = 690

Min number of Atoms per Keyphrase = 1

onmogul.com

Atom-Keyphrases relationships

Mean number of Keyphrases per Atom = 1.69036455056

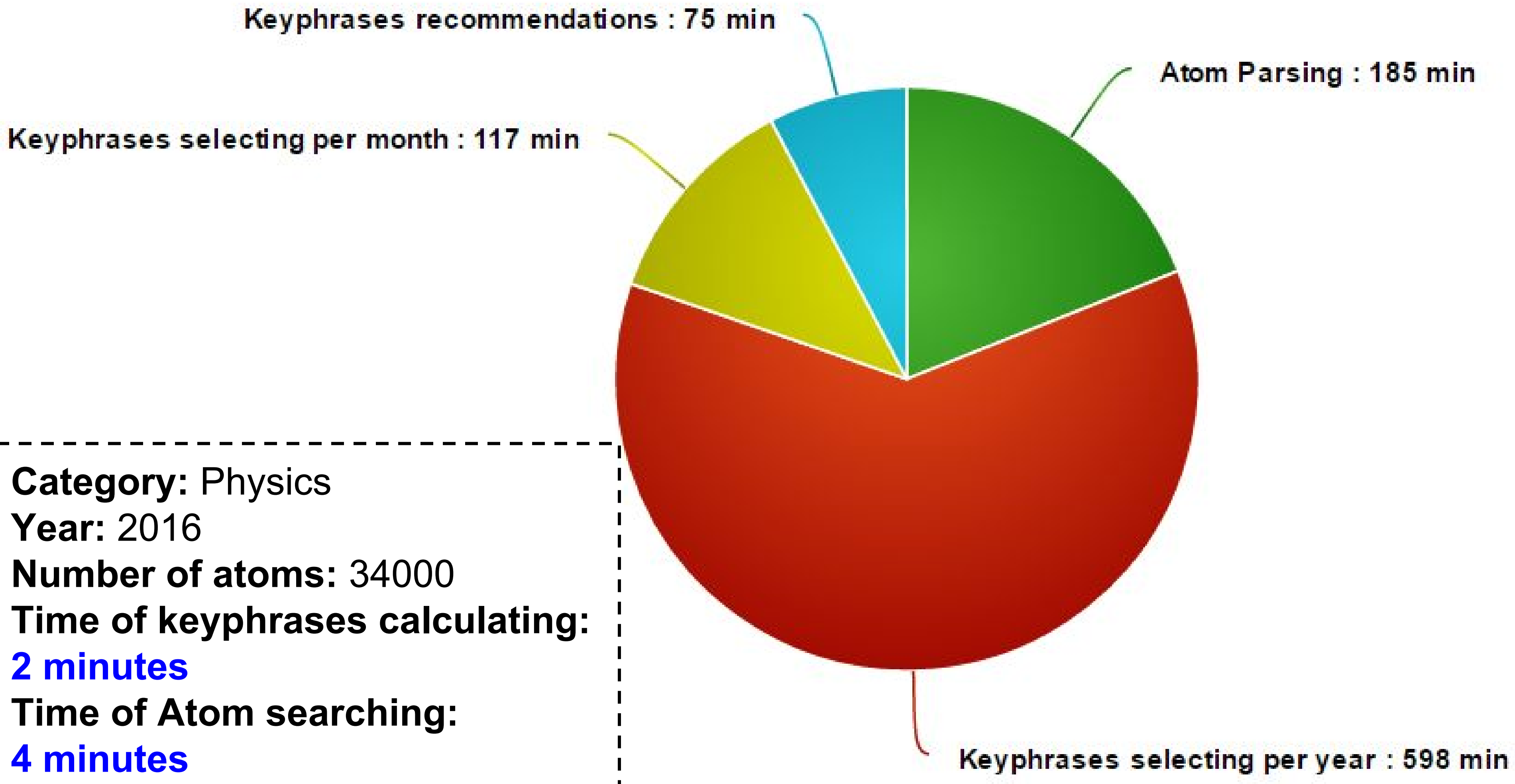
Max number of Keyphrases per Atom = 36

Min number of Keyphrases per Atom = 0

onmogul.com

Processing time

The duration of the general process is about 16 hours



Category: Physics
Year: 2016
Number of atoms: 34000
Time of keyphrases calculating:
2 minutes
Time of Atom searching:
4 minutes
Time of results saving:
2 minutes

AMD FX(tm)-6100 Six-Core Processor
RAM16GB

Components

- **Harvester** responsible for keyphrases extraction
- **Visualization** responsible for application
- **MySQL** is used as a storage

Harvester tools

NLTK

MySQL

scikit learn

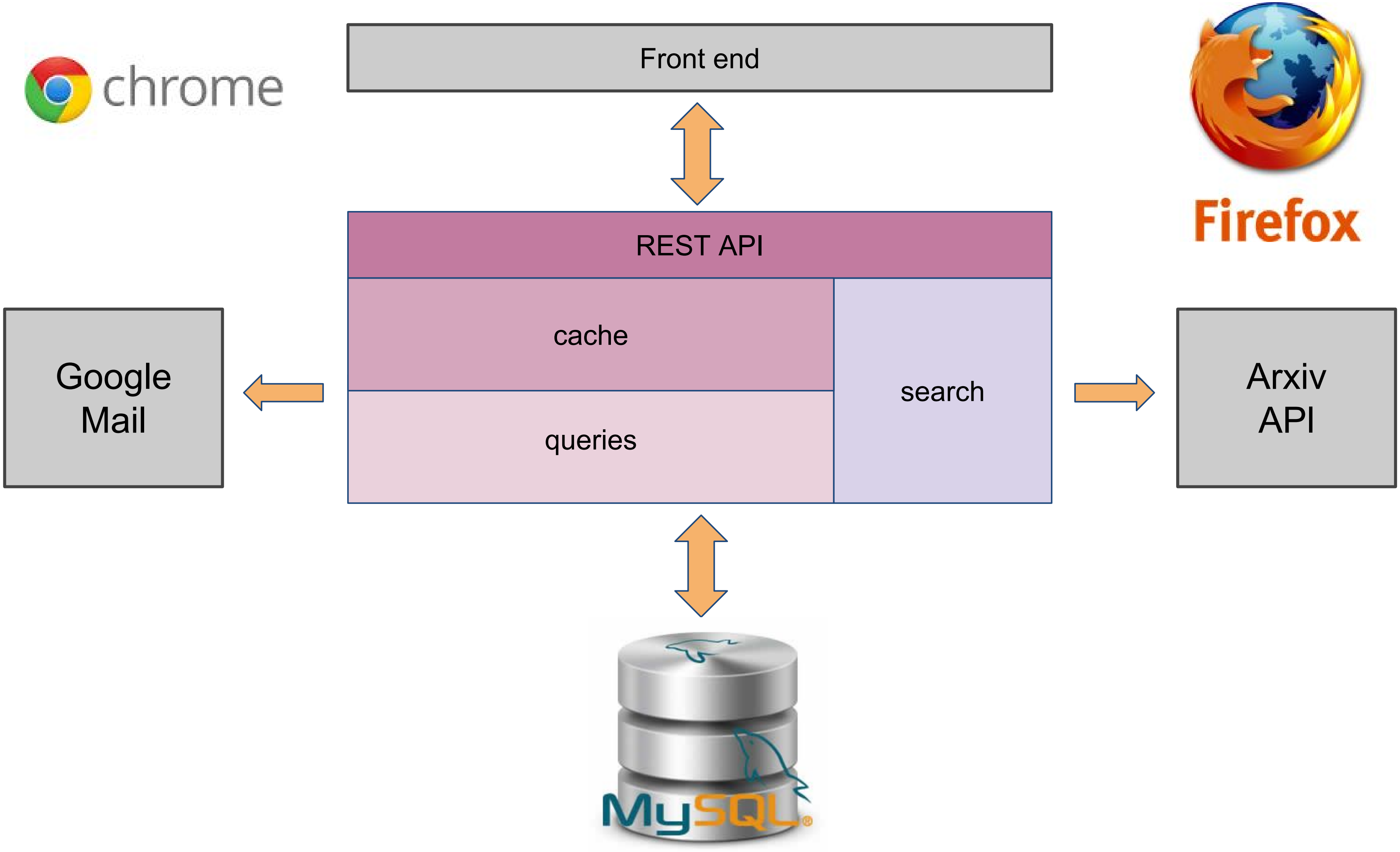
NumPy

XML

gensim

urllib2

Visualization architecture



Visualization stack

ANGULARJS by Google

SVG HTML CSS JavaScript

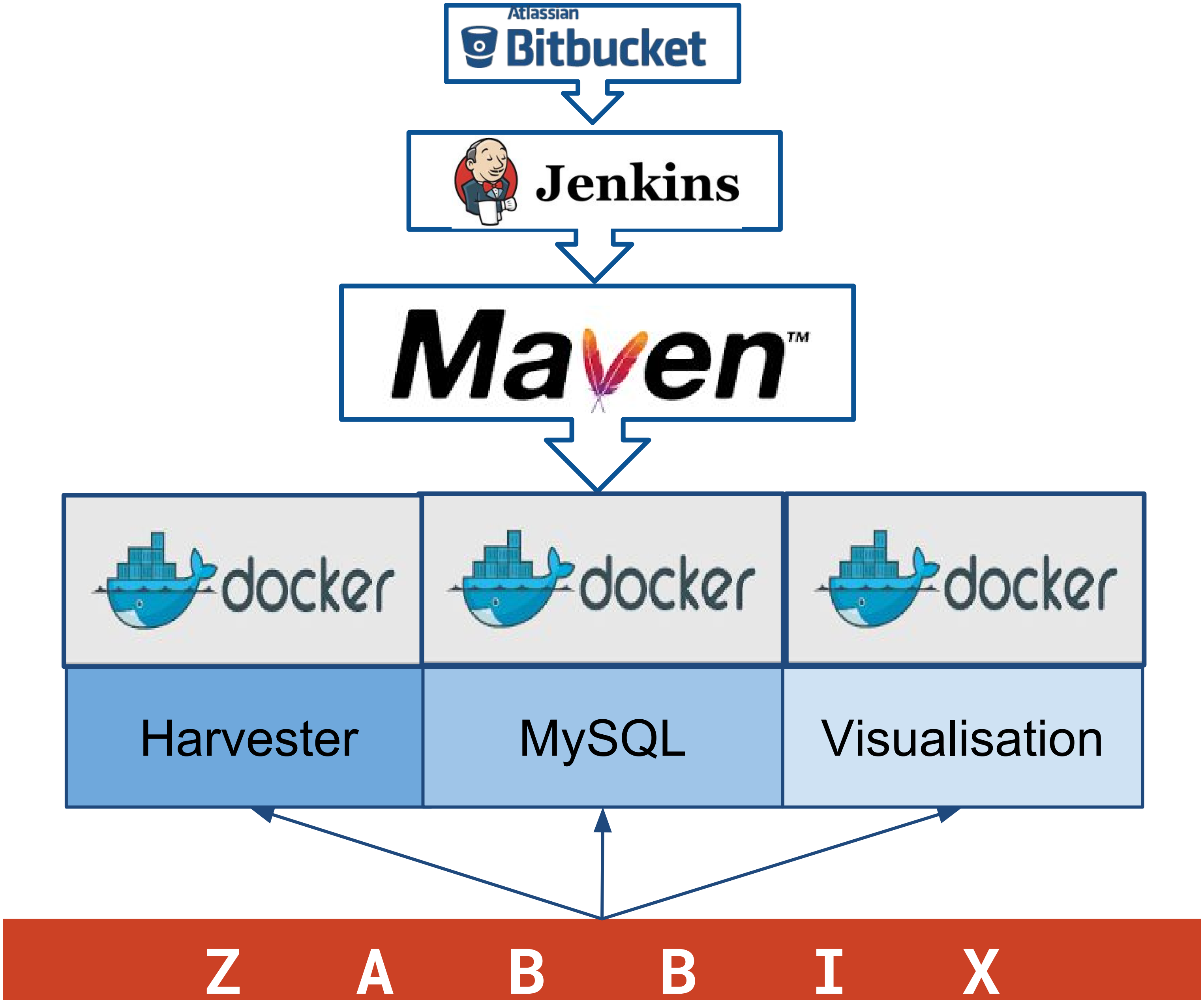
D3.JS - Data Driven Documents

Scala

Bootstrap

spring boot

Deployment pipeline



Harvester

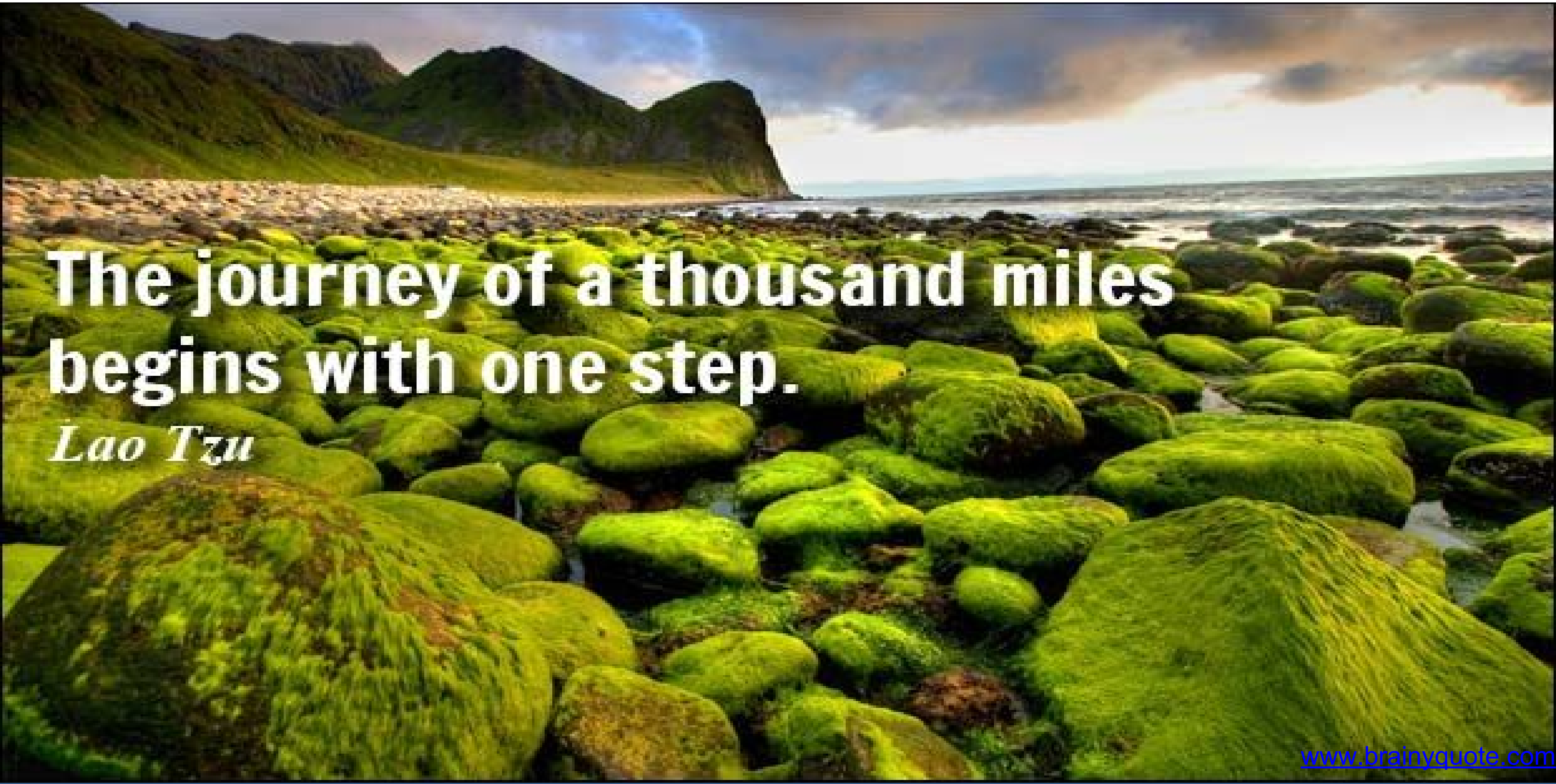
MySQL

Visualisation

ZABBIX

The best V.I.Tech team

SUMMARY



www.brainyquote.com

Plans for future

- Add trends, which people search
- Add trends extraction per subcategory
- Add trends analysis of other sources
- Add Author’s analysis, H-index calculating
- Add Google Analytics
- Scoring

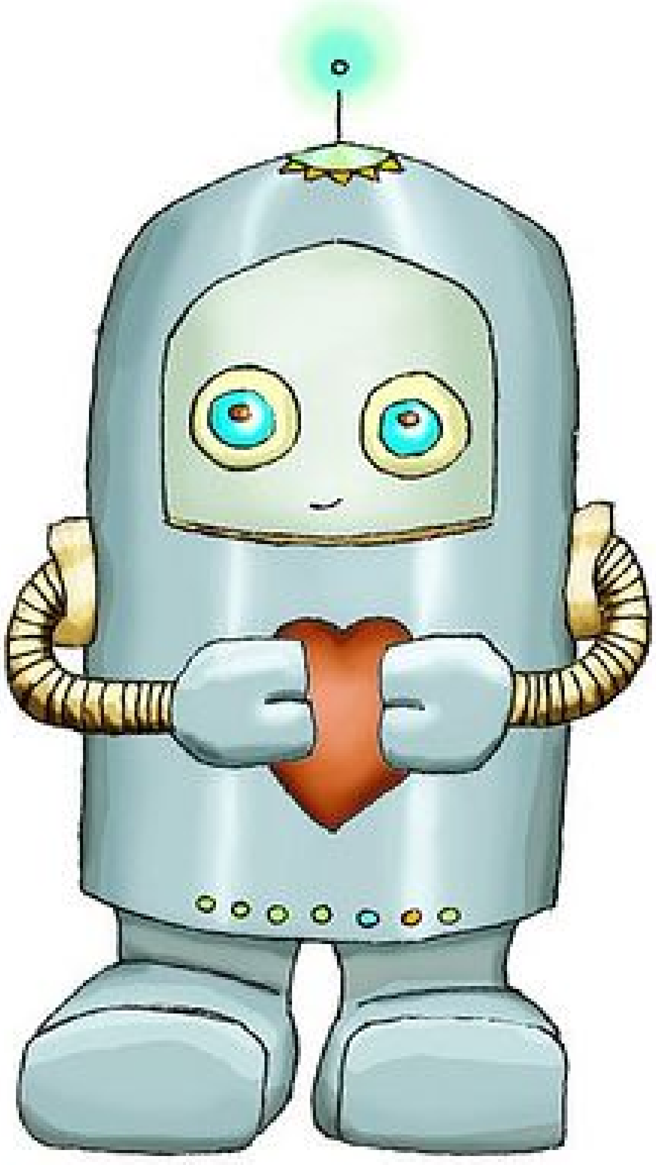
blazepress.com

What keyphrases could you extract from our talk?

Keyphrases from our talk by SciencePulse

Keyphrases	Weight
natural language processing keywords extraction	1.0
method rapid automatic keyword extraction	1.0
scientific organizations explore artificial intelligence	0.7
scientists e-print repository arXiv	0.7
extracting hot topics	0.67
economize scientists time	0.67
human product generally text data	0.67

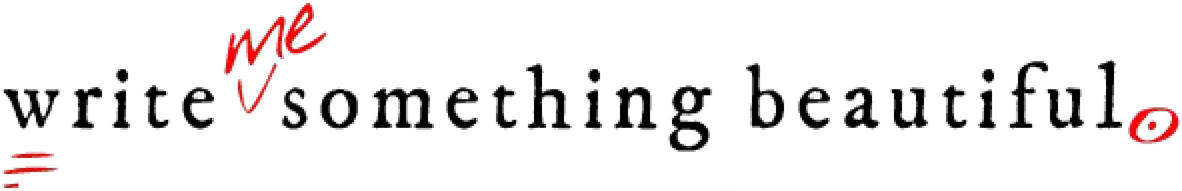
Thank you!



Useful links

http://sciencepulse.vitech.com.ua/

http://ijarcsse.com/docs/papers/Volume_6/5_May2016/V6I5-0392.pdf
https://web.eecs.umich.edu/~mihalcea/papers/mihalcea.emnlp04.pdf
https://hassetukda.wordpress.com/2012/09/24/ukda-keyword-indexing-with-a-skos-version-of-hasset-thesaurus/



tetiana.kodliuk@vitech.com.ua