Natural language processing and the intelligent machines

Osvaldo N. Oliveira Jr

chu@ifsc.usp.br

NILC

São Carlos Institute of Physics

University of São Paulo, Brazil

NILC

www.nilc.icmc.usp.br



















Outline

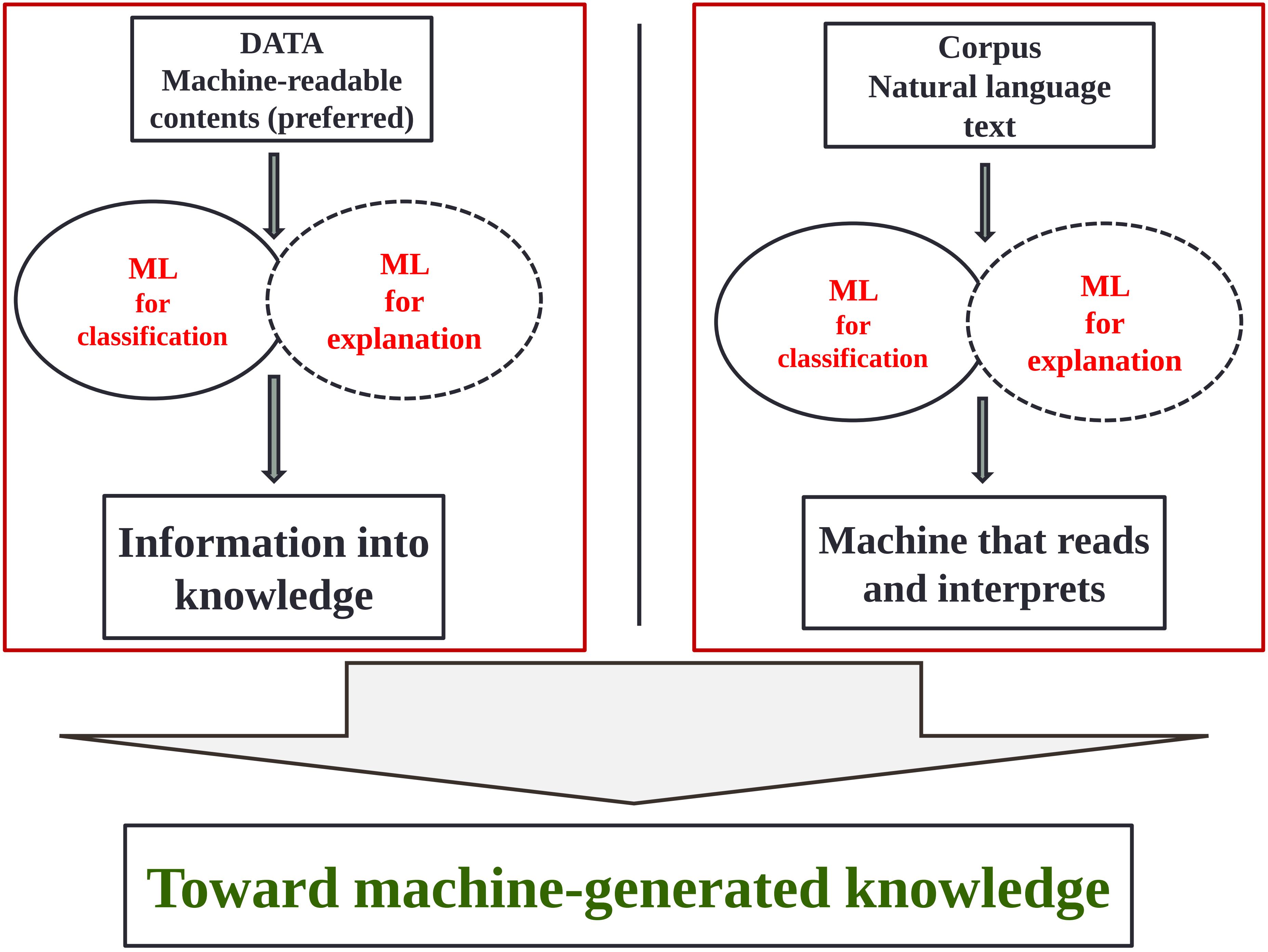
- The Fifth Paradigm
- Big Data and Machine Learning
- Text analytics and computer-assisted diagnosis
- Complex networks for NLP

The Fifth Paradigm

- 1st Empirical, descriptive
- 2nd Theory and experiment
- 3rd Theory, experiment, computer simulation
- 4th All of the above + Big Data

The Fifth Paradigm

- 1st Empirical, descriptive
- 2nd Theory and experiment
- 3rd Theory, experiment, computer simulation
- 4th All of the above + Big Data
- <u>5th Machine-generated knowledge</u>


DATA
Machine-readable
contents (preferred)
ML
for
classification
ML
for
explanation
Information into
knowledge
Corpus
Natural language
text
ML
for
classification
ML
for
explanation
Machine that reads
and interprets
Toward machine-generated knowledge

Some Requirements

- Text analytics – large text databases
- Lots of data: experimental, theoretical (DFT, etc) and simulation (MD, etc)
- Internet of Things
- Machine Learning Methods (Deep Learning, etc)

Computer-assisted diagnosis as an example

My talks over a decade or so

Limitations from hardware X software

Ambiguities

Syntax, semantics, pragmatics

Number 10 has congratulated the Reds. They are now as famous as their Penny Lane fellows.

Need of understanding local culture as example of the difficulties

Computer vs. Human

How IBM's Watson Computer Excels at *Jeopardy!* By John Rennie

NLP + Machine Learning

Hybrid approach – symbolic + corpus-based

Watson occupies the space of 10 refrigerators, with 90 servers having 3290 processors each. It may process 500 GB/s, corresponding to 1 million books.

Each server has 256 GB of RAM, and may store 200 million pages. In Jeopardy, hard disks are not used, for access would be slow. A lot of parallel processing

How are the questions interpreted?

Speech processing is not necessary, for Watson employs the text provided.

Methodology: apparently a mix of strategies from traditional Q&A (questions-answers) systems together with machine learning from examples.
Watson also estimates the probality of having the right answer.

Todai Robot Project

Noriko Arai, National Institute of Informatics, Japan

AI system that answers real questions of university entrance exams consisting of two parts, the multiple-choice style national standardized tests and the written tests including short essays.

From 2013, the software has taken mock tests of the National Center Test every year.

Top 1% in Maths

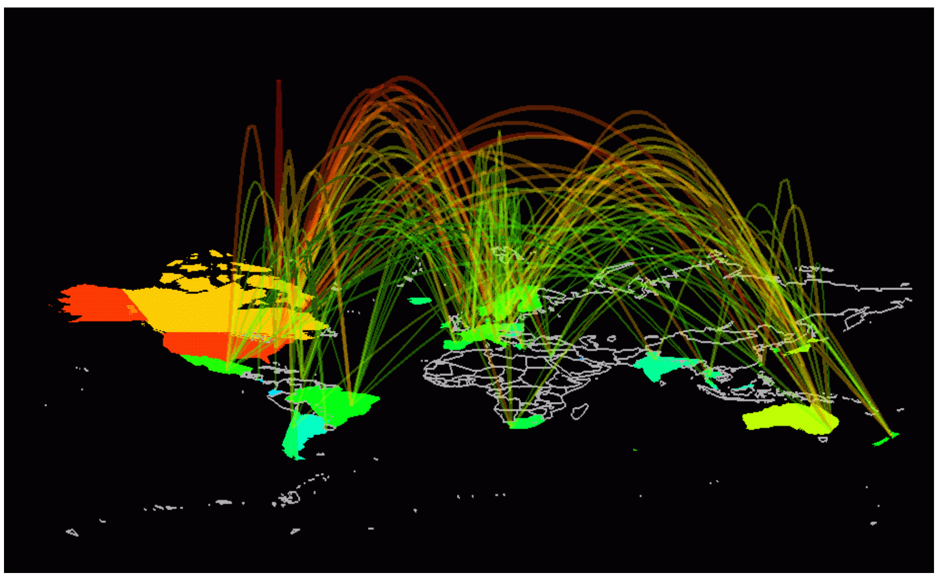
Its ability is still far below the average entrants of Tokyo University. However, it is beyond the average: it is competent to pass the entrance exams of two thirds of universities including 33 national universities in Japan.

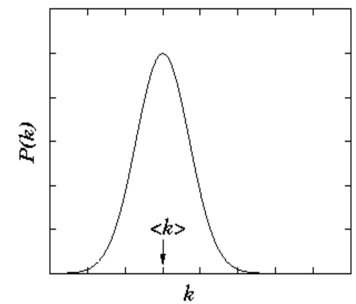
Ubiquity of classification tasks

Metrics and measurements

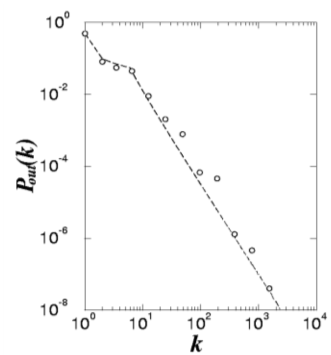
- **Metrics from first-order statistics**
- **Metrics from networks representing the system**
- **Time series extracted from the system**

Internet Backbone





Expected



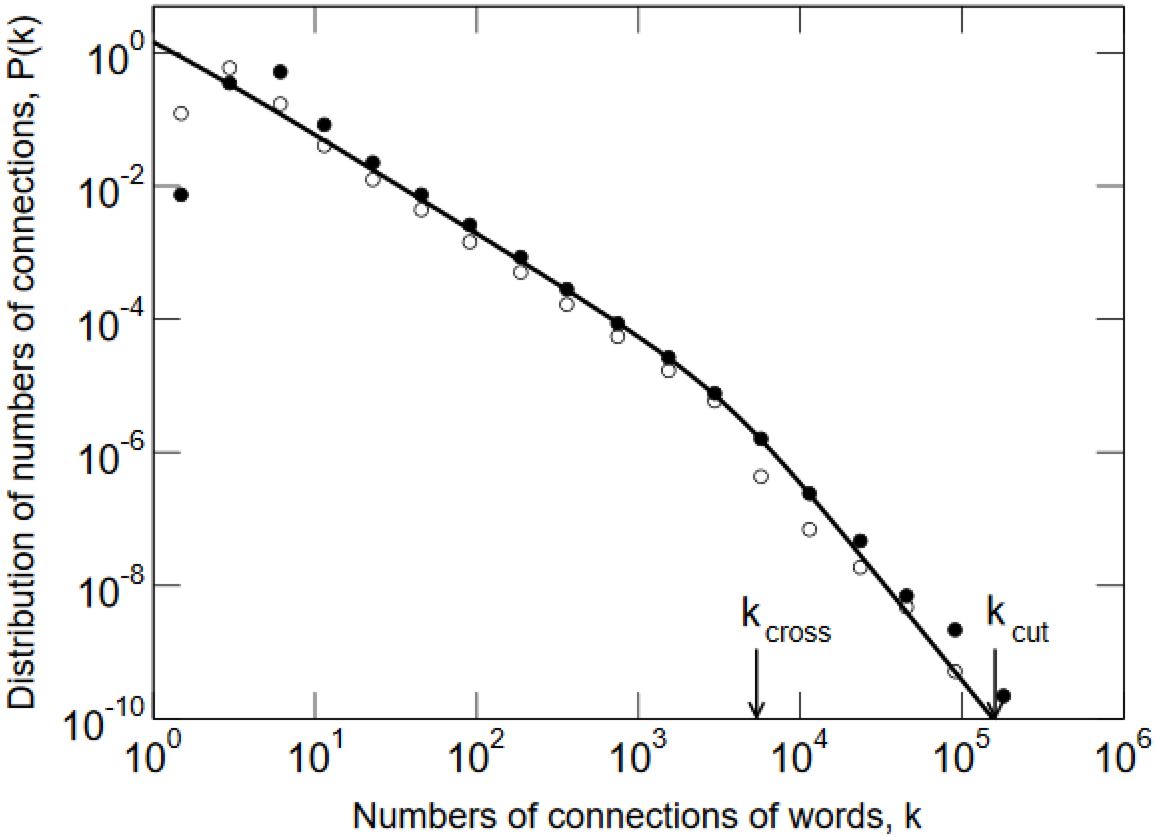
Found

Barabási, Sci. American, 2003

Scale-free network

Applications of Complex networks, Costa et al., Adv. Phys., 2011

Text as complex network



Web of words is a scale-free network

Dorogovtsev and Mendes, *Advances in Physics*, 2002

Processing text

Pre-processing and data acquisition

- **Identify and remove stopwords**
- **Lemmatization**
- **Dealing with punctuation and paragraphs**
- **Obtaining statistical measurements (usually frequency related)**
- **Creating co-occurrence networks**
- **Analysis at the word and then text levels**

Source text represented with nodes connected by edges

nodes → *text elements*

(e.g., words, sentences)

edges → *linguistic relations*

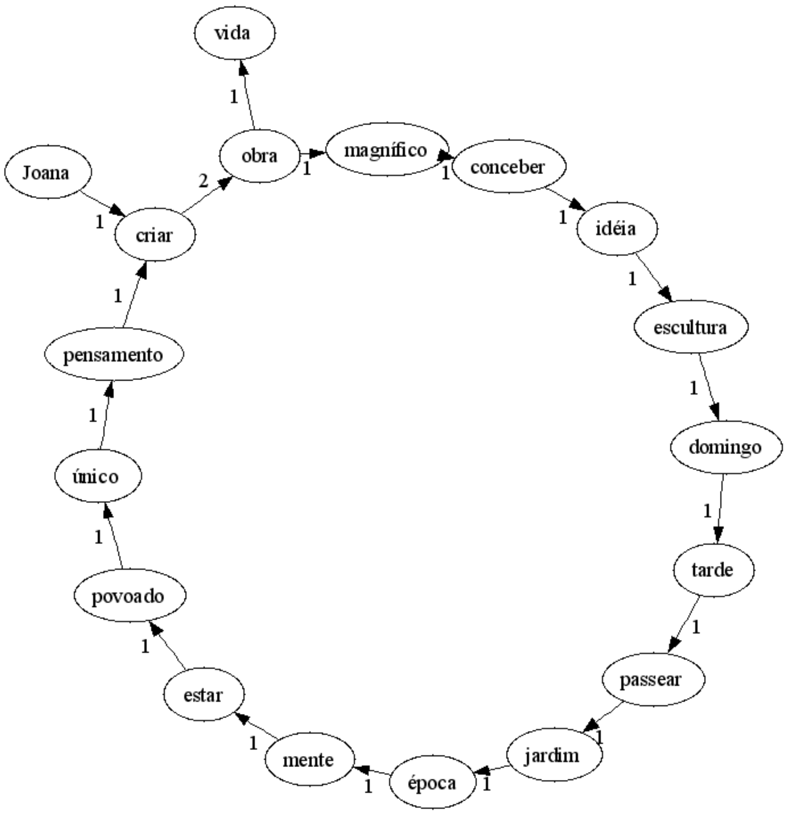
(e.g., syntactic, semantic, co-occurrence)

edges → *type/strength of relations*

(e.g., word frequency)

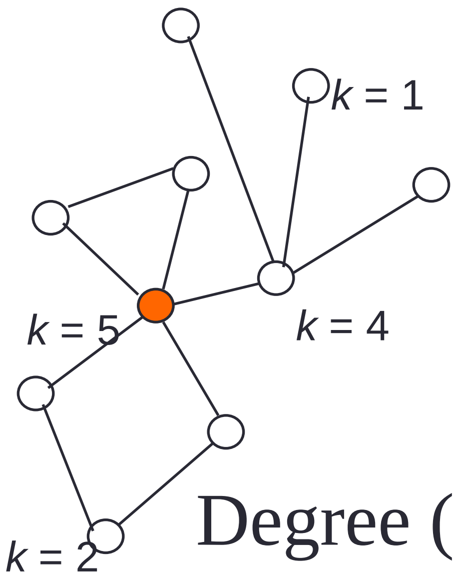
Text as complex network

Joana criou uma obra magnífica. Concebeu a idéia de sua escultura num domingo à tarde, ao passear pelo jardim. Nessa época, sua mente estava povoada por um único pensamento: criar a obra de sua vida.



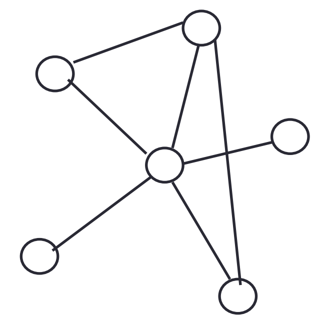
Joana criar uma obra magnífico. Conceber a idéia de sua escultura num domingo à tarde, ao passear pelo jardim. Nessa época, sua mente estar povoado por um único pensamento: criar a obra de sua vida.

Treating systems as networks



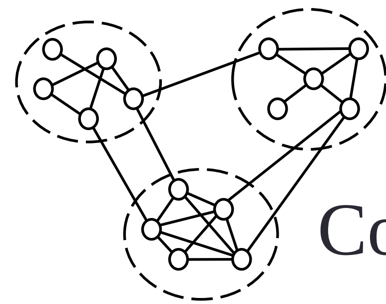
Degree (k)
and hubs

Cluster coefficient



$$C = \frac{e_c}{e_T} = \frac{2e_c}{(n)(n-1)}$$

$$= \frac{(2)(2)}{(5)(5-1)} = 0.2$$



Communities

Other metrics:

Distances, shortest paths, communities, borders, accessibility, hierarchical degrees, centrality, node activity ..

More than 100 have been used for topology and dynamics

Summarization

Pardo et al., Modeling and evaluating summaries using complex networks, Lecture Notes in Artificial Intelligence, 2006.

Antiqueira et al; A complex network approach to text summarization; Information Sciences, 2009.

Amancio et al.; Extractive summarization using complex networks and syntactic dependency, Physica A, 2012.

Evaluation of Machine Translation

Amancio et al.; Complex networks analysis of manual and machine translations, International J. Mod. Phys. C, 2008.

Amancio et al.; Using metrics from complex networks to evaluate machine translation, Physica A, 2011.

The Voynich

Amancio et al;; Probing the Statistical Properties of Unknown Texts: Application to the Voynich Manuscript, PLOS One, 2013.

Language analysis – including text quality evaluation

Antiqueira et al.; Strong correlations between text quality and complex networks features, Physica A, 2007.

Amancio et al.; Using complex networks to quantify consistency in the use of words, J. Stat. Mech., 2012.

Amancio et al.; Complex networks analysis of language complexity, Eur. Phys. Lett., 2012.

Amancio et al.; Structure–semantics interplay in complex networks and its effects on the predictability of similarity in texts, Physica A, 2012.

Amancio et al; Unveiling the relationship between complex networks metrics and word senses, Europhys. Lett., 2012.

Language as sensors

Dos Santos et al.; Enriching Complex Networks with Word Embeddings for Detecting Mild Cognitive Impairment from Speech Transcripts, ACL, 2017

Scientometrics

Amancio et al.; Three-feature model to reproduce the topology of citation networks and the effects from authors' visibility on their h-index, J. Informetrics, 2012.

Amancio et al.; On the use of topological features and hierarchical characterization for disambiguating names in collaborative networks, Eur. Phys. Lett., 2012.

Amancio et al.; Using complex networks concepts to assess approaches for citations in scientific papers, Scientometrics, 2012.

Silva et al.; Quantifying the interdisciplinarity of scientific journals and fields, J. Informetrics, 2013.

Amancio et al.; Topological-collaborative approach for disambiguating authors' names in collaborative networks, Scientometrics, 2015.

Authorship Identification

Amancio et al., Comparing intermittency and network measurements of words and their dependence on authorship, New J. Phys., 2011.

Amancio et al.; Identification of literary movements using complex networks to represent texts, New J. Phys., 2012.

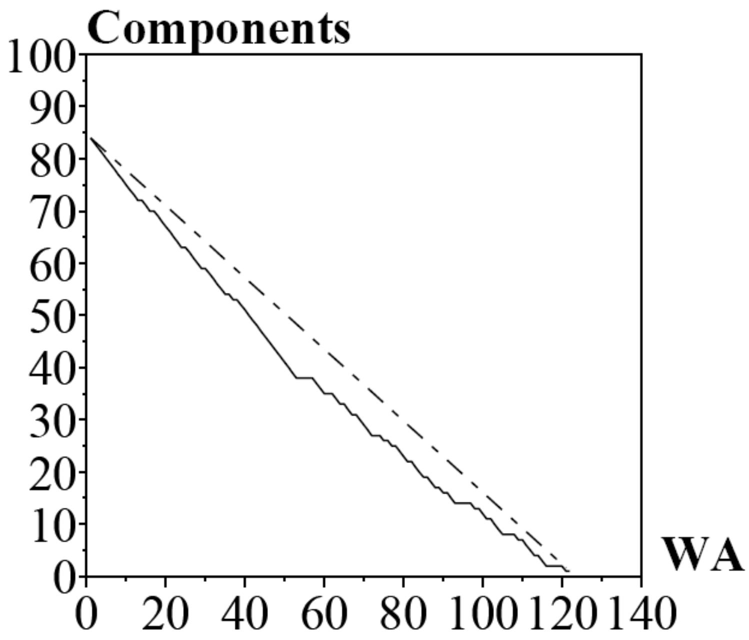
Akimushkin et al.; Text Authorship Identified Using the Dynamics of Word Co-Occurrence Networks, PLOS One, 2017.

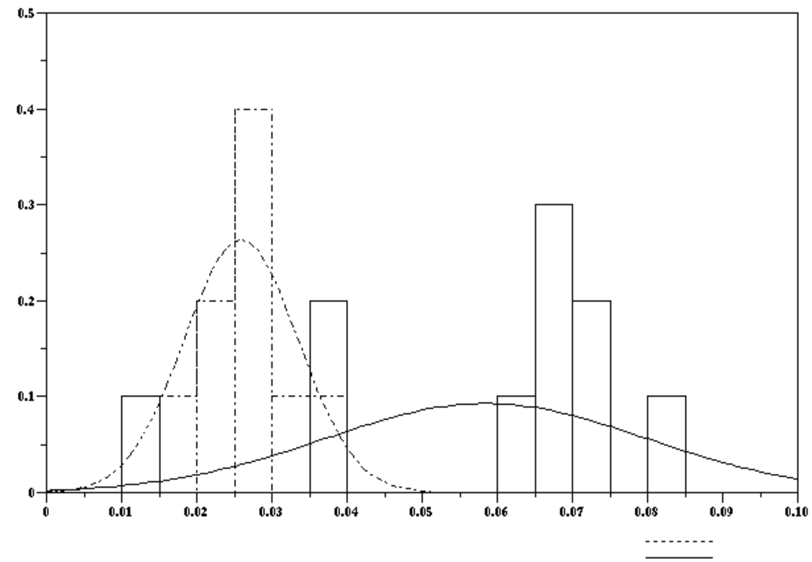
Akimushkin et al.; On the role of words in the network structure of texts: application to authorship attribution, Physica A, 2018.

Semi-automated Surveys

Silva et al.; Using network science and text analytics to produce surveys in a scientific topic, J. Informetrics, 2016.

Dynamics of a network

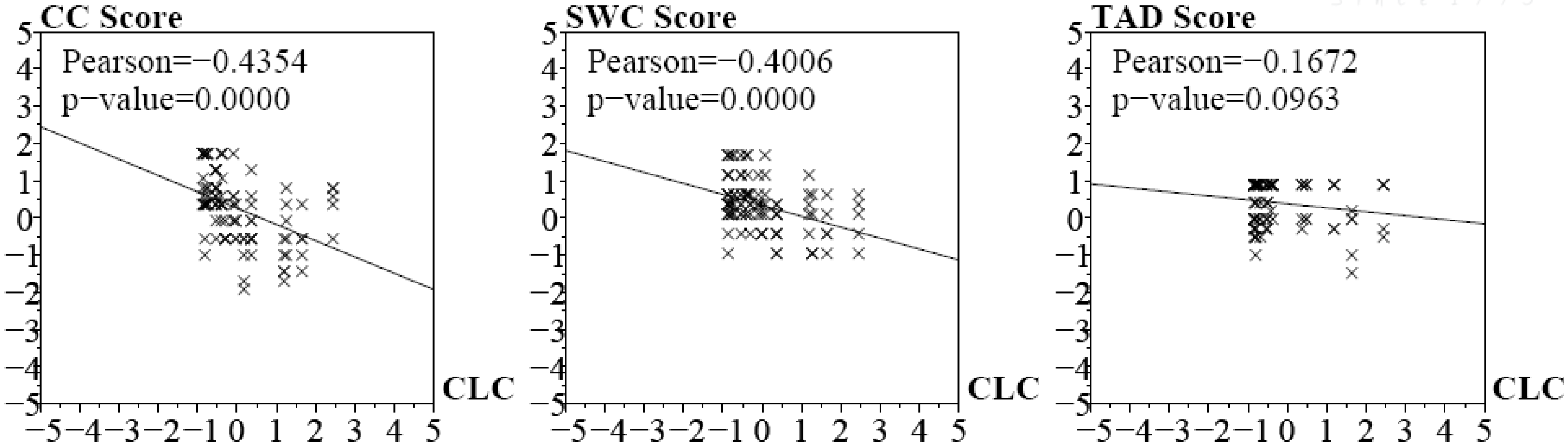




Dynamics may be indicative of text quality

Antiqueira et al., *Physica A*, 2007

Network features

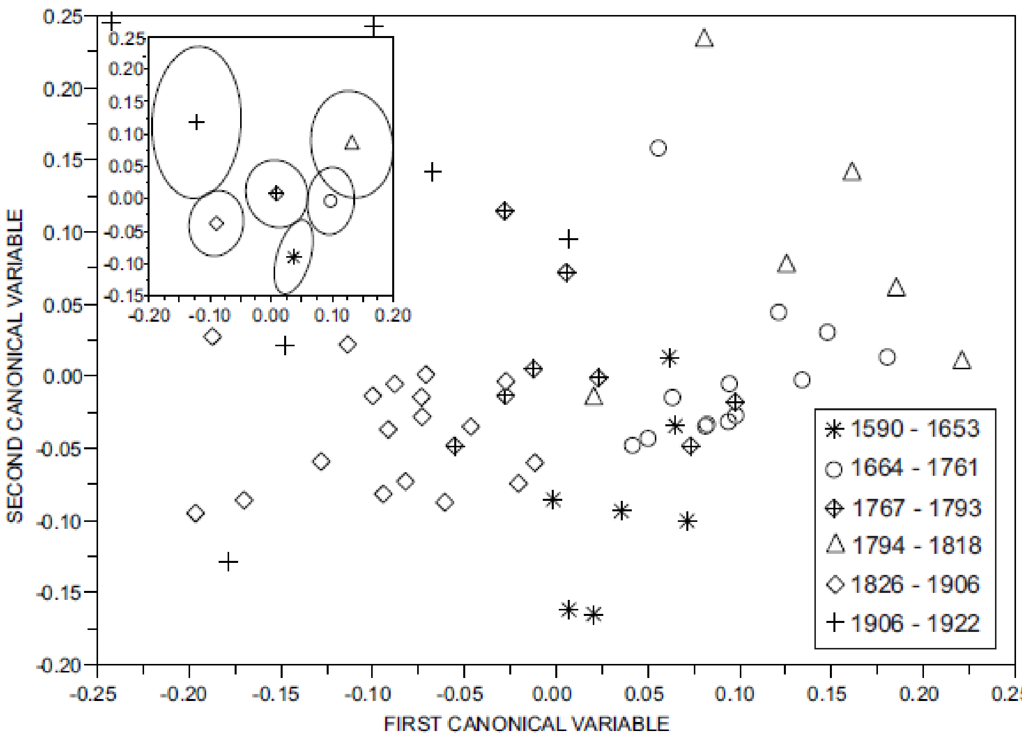


Human evaluation with regard to coherence and cohesion (CC), standard writing convention (SWC) and topic adherence (TAD) correlates well with the clustering coefficient

Antiqueira et al., *Physica A*, 2007

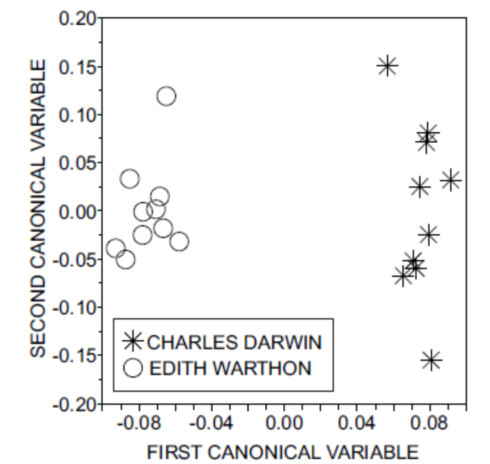
Relationship between the best clustering of writing styles the traditional classification of literary movements.

Cluster Boundary	Literary Boundary	Literary Movement
1590 – 1653	1558 – 1903	Elizabethan era
1664 – 1761	1660 – 1798	Neoclassicism/ Enlightenment
1767 – 1793	1660 – 1798	Neoclassicism/ Enlightenment
1794 – 1818	1764 – 1820	Gothic fiction
1826 – 1906	1830 – 1900	Realism
1826 – 1906	1865 – 1900	Naturalism
1906 - 1922	1890 - 1940	Modernism



Literary Movements

Identification of movements using complex networks to represent text

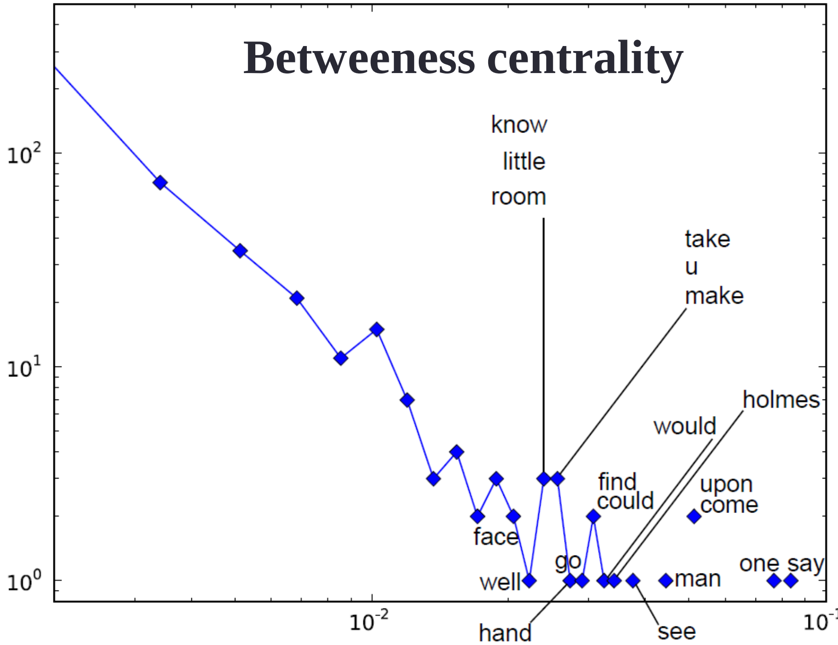


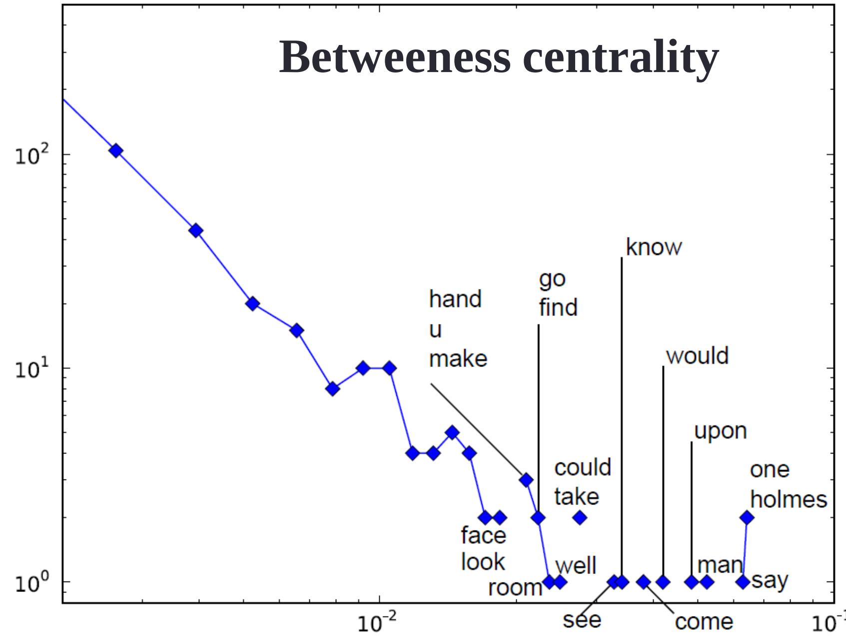
Darwin vs Edith Warthon

Amancio et al, *New J. Phys. 2012*

Role of words in network structure

Dissimilarity matrices for 4 metrics (degree, shortest paths, betweeness, intermittency). Only for 100 most relevant nodes



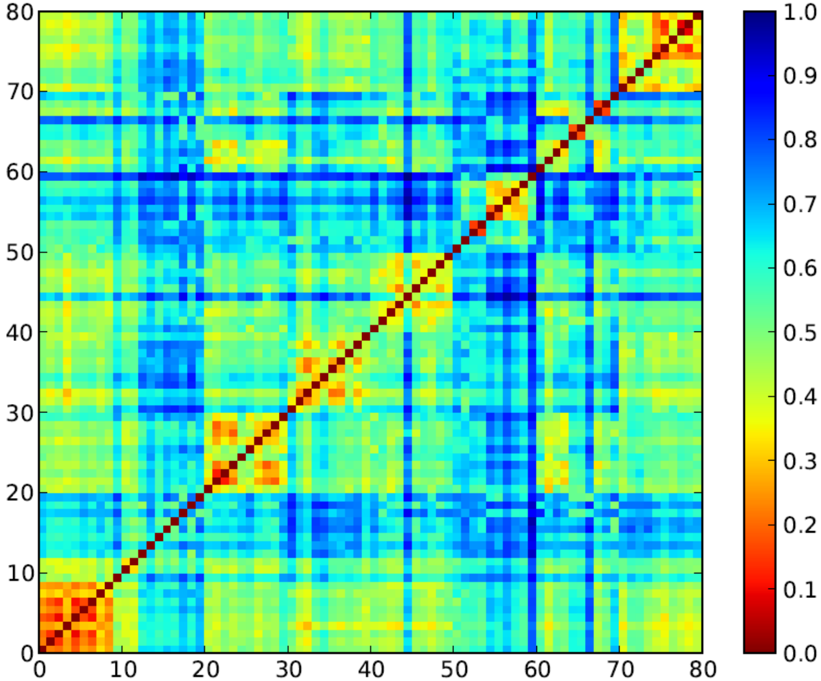


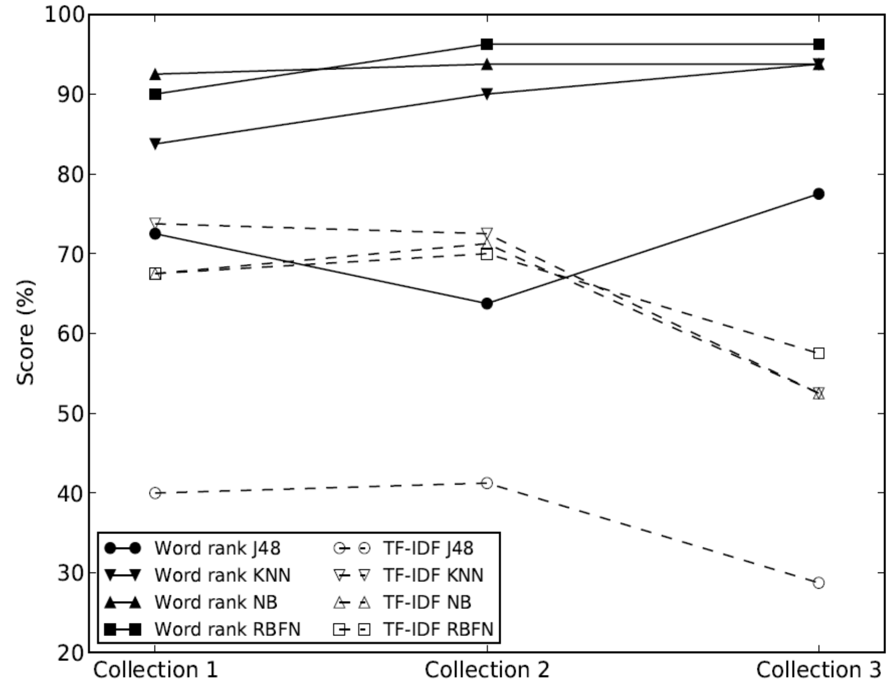
Two Sherlock Holmes novels: Only one different word among the first 20

Similarity (dot product) between 2 texts is high if the same words occupy similar positions in the distributions

Akimushkin et al., Physica A., 2018

Efficient authorship attribution



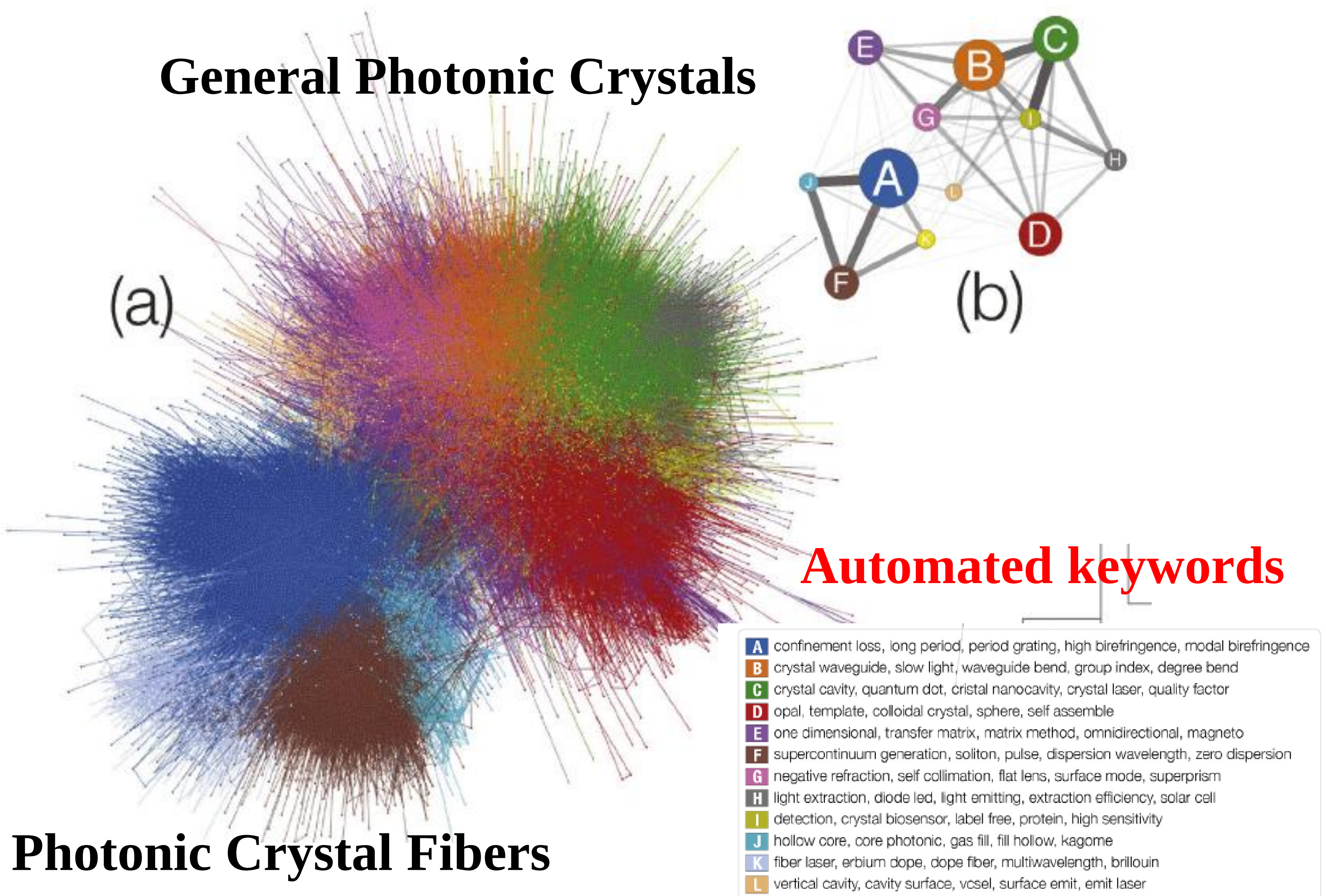


Dissimilarity matrices with multi-dimensional scaling for feature selection. Radial Basis Function Network (RBFN) yields the highest scores

Akimushkin et al., Physica A., 2018

Semi-automated Surveys

General Photonic Crystals

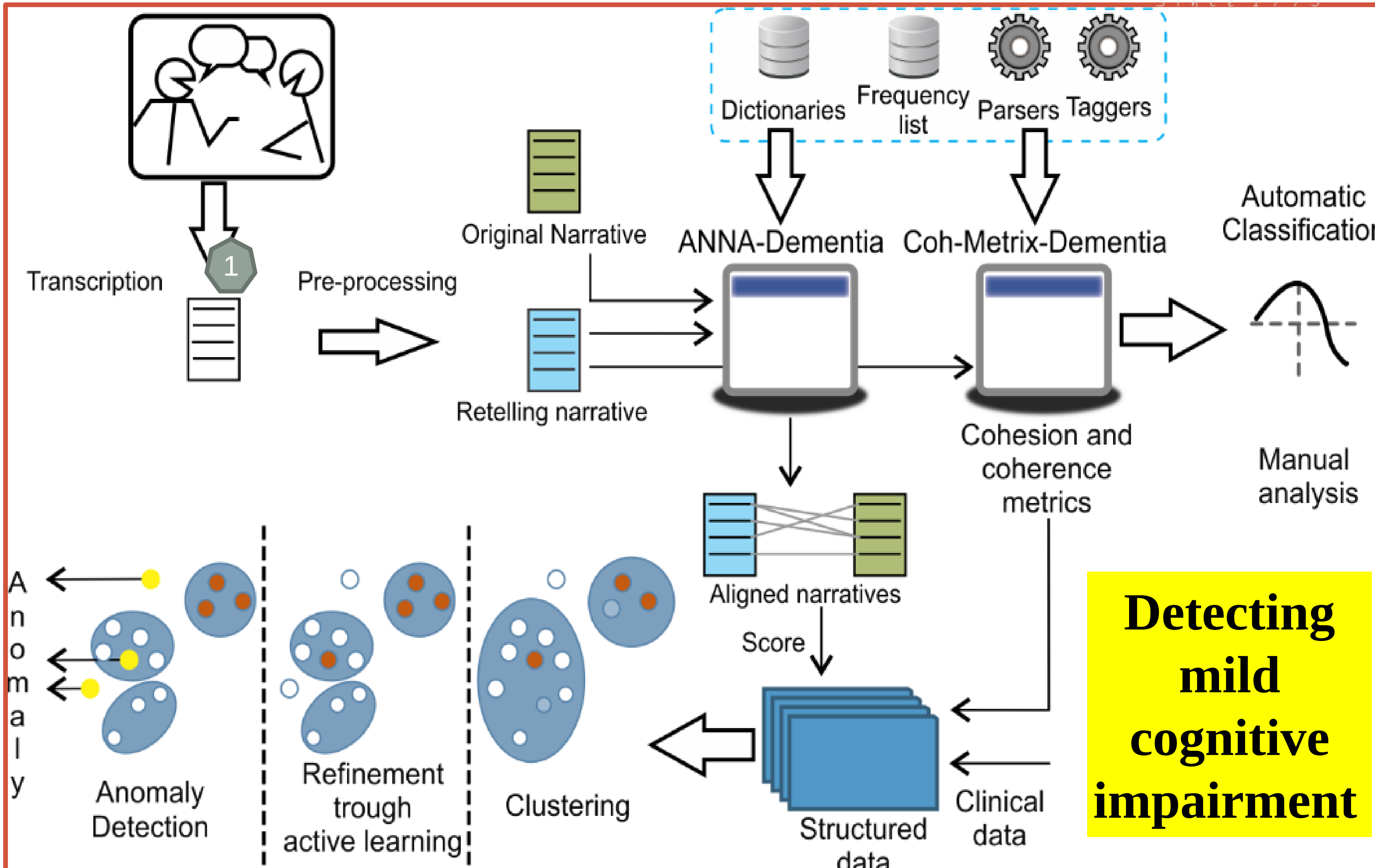


Automated keywords

Photonic Crystal Fibers

F. N. Silva et al, *Journal of Informetrics, 2016.*

Language as sensors

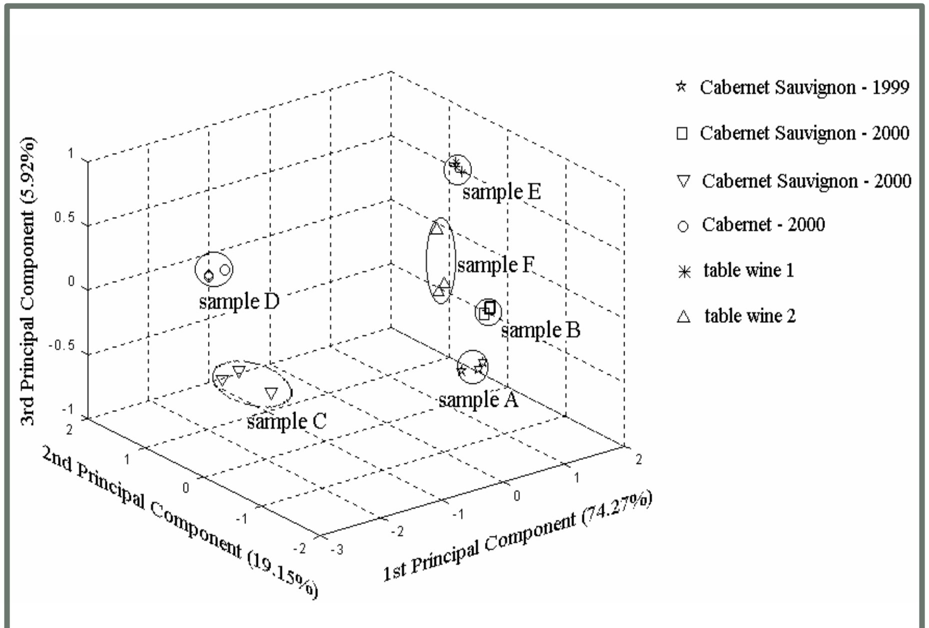


Detecting mild cognitive impairment

With Sandra Aluísio, Letícia Mansur and Diego Amancio, ACL 2017

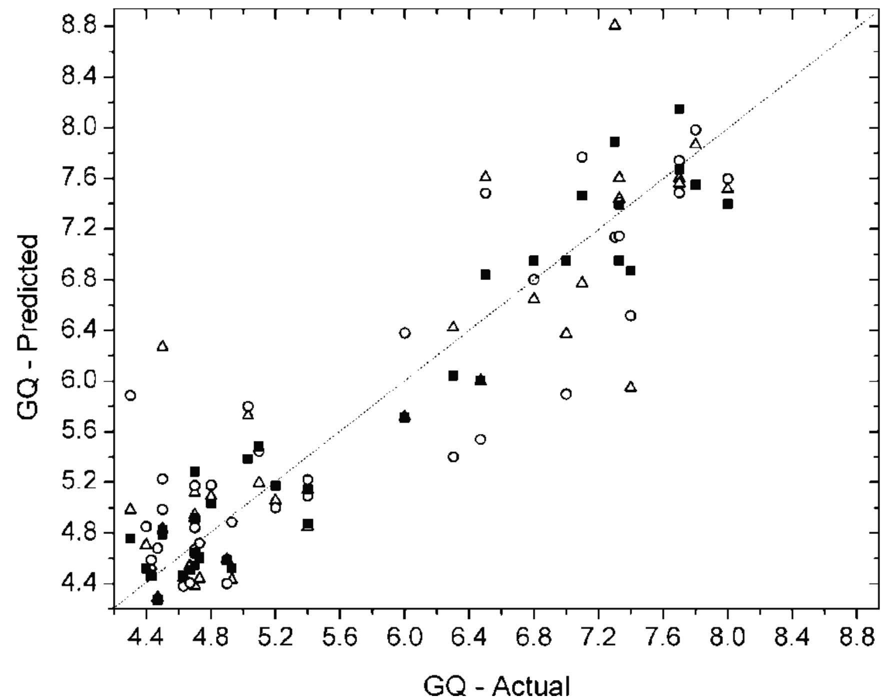
Taste sensors

Identifying wines



Dos Santos Jr. et al.,
Macromol. Biosci., 2003

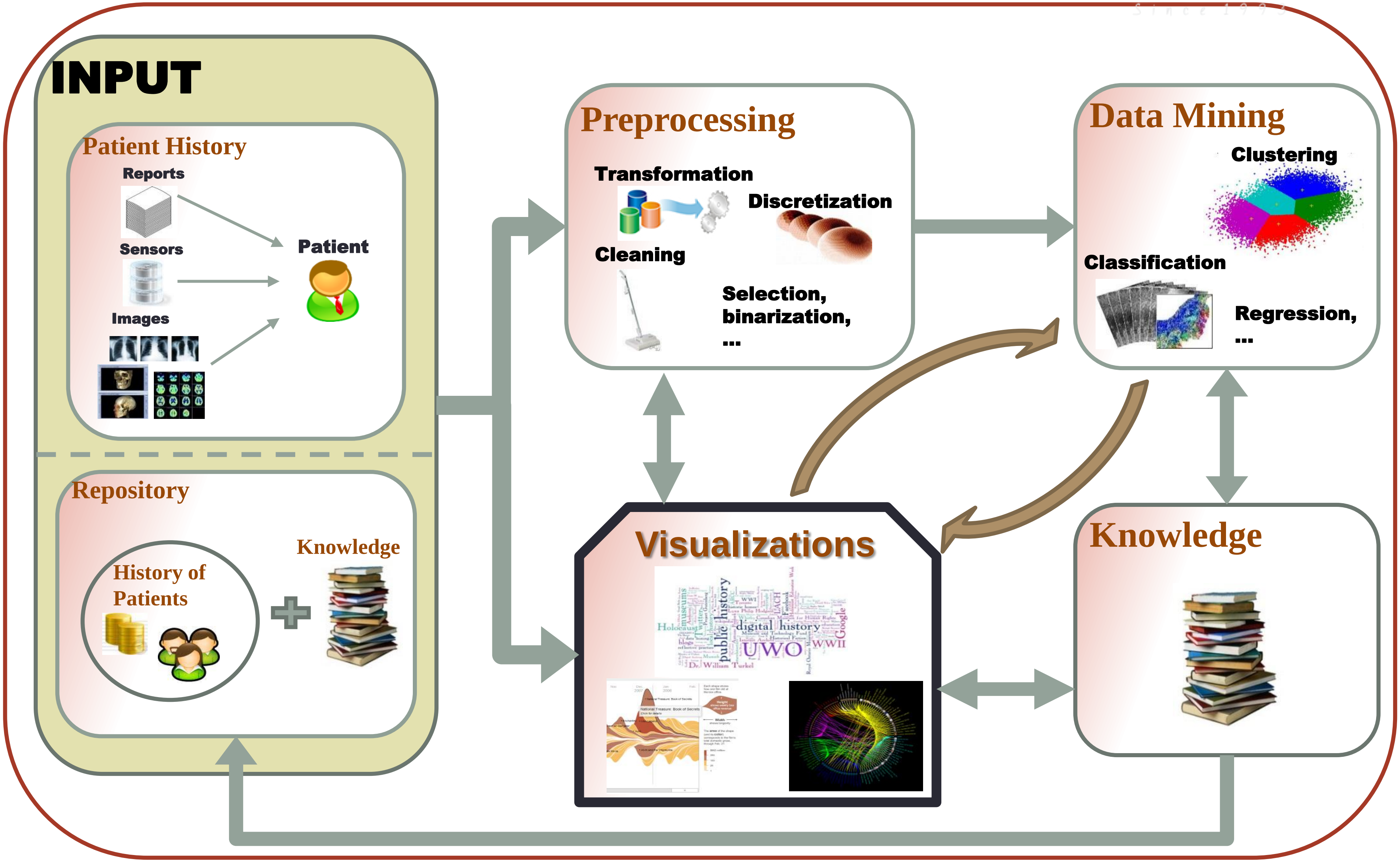
Correlating with human taste



One of the regression methods led to Pearson coefficient of 0.964. Accuracy in the score ± 0.3

E.J. Ferreira et al.,
Electronics Letters, 2007

Holy Grail: Diagnostics in the future



Oliveira et al., Chem. Lett. Japan, 2014

Disclaimers

- **Present approaches cannot replace refined analysis (human). Yet!**
- **In most cases only correlations can be established**
- **Often, reasonable explanations are not to be found to account for the findings**

Interpretation in ML

Could an interpretation task consist of classification subtasks?

From a lecture in Scientific Writing

Getting the message across

- **What are the contributions of your work?**
- **Why are these contributions important?**

Organizing ideas and results

- **What are your key findings?**
- **What is the importance of your key findings to the field?**
- **Are your findings complete? If not, what is missing?**
- **What is the supporting evidence?**
- **Do they provide a basis for publication?**

A vision for a survey

- Determine the structure and ontology of a research topic (actual surveys)
- Use ML to teach how the survey should be organized
- Develop a specific Q&A system for the questions posed about an article

Final Remarks

- NLP is becoming ubiquitous and is one of the most important fields for science, technology and society
- ML-based methods are (or will) dominating
- NLP cannot be isolated from other fields. It is too central for that to happen

The machines of the future

Machine learning will change the landscape of science and technology in the XXI century.

In a few decades, most intellectual tasks will be better performed by machines.

Is society being prepared for that?

Final Recommendation/Provocation

- How would an intelligent machine solve the scientific problem you are addressing?
- Are you sure the problem could not be obviated by other means?

Acknowledgments

Luciano F. Costa, Diego R. Amancio, Graça Nunes, Sandra M. Aluísio, Thiago Pardo, Maria Cristina F. de Oliveira, Lucas Antiqueira, Camilo Akimushkin, Renato Fabbri, Vilson Vieira, Filipi N. Silva, Aretha Alencar, Rosane Minghim, Fernando Paulovich, Francisco A. Rodrigues, José Fernando Rodrigues Jr., Gonzalo Travieso, Rafael M. Maki, Leandro B. dos Santos, Edilson A. Corrêa Jr., Letícia L. Mansur, Paulino Vilas Boas (USP), Antonio Pessotti, Eleonora Albano (Unicamp), Eduardo Altmann – Univ. Sydney (Australia), Maria Bardosova (Cork, Ireland)

Computers today

- Same for all
- Hundreds to thousands of programs and apps (only)
- Slow, manual learning
- No independence to make decisions

Computers in 20, 30 years

- Customized
- Millions of programs and apps
- Heritage will be relevant.
  Machine educated permanently
- Impossible to determine who
  make decisions

Apps

- Distributed, collaborative development
- Focused on problem solving
- Multidisciplinary
- Take advantage of IT and Big Data infrastructure

In consonance with all drivers for change and requirements for professional training

WRITING SCIENTIFIC PAPERS IN ENGLISH SUCCESSFULLY

YOUR COMPLETE ROADMAP

ETHEL SCHUSTER | HAIM LEVKOWITZ | OSVALDO N OLIVEIRA JR.
(EDITORS)

SCHUSTER | LEVKOWITZ | OLIVEIRA JR. (EDITORS)
WRITING SCIENTIFIC PAPERS IN ENGLISH SUCCESSFULLY

SANDRA MARIA ALUÍSIO
is a lecturer in computer science at the Sao Carlos Institute of Mathematics and Computer Science at the University of Sao Paulo, Brazil, and a member of the Interinstitutional Center for Research and Development in Computational Linguistics (NILC).

CARMEN DAYRELL
currently works as a senior research associate at the ERSC Centre for Corpus Approaches to Social Science at Lancaster University in the United Kingdom.

VALÉRIA DELISANDRA FELTRIM
is a lecturer in computer science in the Informatics Department at the State University of Maringa, Brazil, and a member of NILC.

HAIM LEVKOWITZ
is a computer science faculty member at the University of Massachusetts Lowell, USA.

OSVALDO N. OLIVEIRA JR.
is a professor at the Sao Carlos Institute of Physics at the University of Sao Paulo, Brazil, and a member of NILC.

ETHEL SCHUSTER
is a professor in the Department of Computer and Information Sciences at Northern Essex Community College, Massachusetts, USA.

STELLA E. O. TAGNIN
is a professor in English language,translation, and corpus linguistics in the Department of Modern Languages at the University of Sao Paulo, Brazil.

VALTENCIR ZUCOLOTTO
is a professor at the Sao Carlos Institute of Physics at the University of Sao Paulo, Brazil.

9 788588 533974

ICMC USP
SÃO CARLOS

DL
UFSCar

DC
Departamento de Computação

unesp
/ FCLar

IFSC

EACH
USP

INSTITUTO
FEDERAL
SÃO PAULO
Campus São Carlos

UNIVAS
UNIVERSIDADE DO VALE DO SAPUCAÍ

Disce Doce
RERVM COGNOSCERE CAVSAS
The University Of Sheffield.

IBILCE
Instituto de Biociências,
Letras e Ciências Exatas

UEM

Embrapa
Informática Agropecuária