SCIS&ISIS2020

December 5-8, 2020
Online Conference

Interactive Natural Language Technology for Human-centric Explainable Artificial Intelligence

Jose M. Alonso

josemaria.alonso.moral@usc.es

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

CAMPUS VIDA CAMPUS DE EXCELENCIA INTERNACIONAL

CiTIUS Centro Singular de Investigación en Tecnoloxías Intelixentes

SCIS-ISIS

What is Intelligence?

Are they intelligent ? What features characterize their intelligence ?

Autonomy

Frames of Mind: the Theory of Multiple Intelligences
(Howard Gardner)

Learning

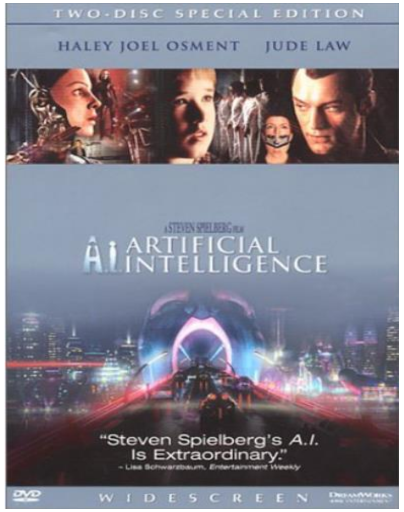
Why ?

Knowledge

Centro Singular de Investigación en Tecnoloxías Intelixentes

Jose M. Alonso
https://citius.usc.es/v/jose-maria-alonso-moral

What is Artificial Intelligence?



✓ The ability of a digital computer or computer-controlled robot to perform **tasks commonly associated with intelligent beings**. The term is frequently applied to the project of developing systems endowed with the intellectual processes characteristic of humans, such as the **ability to reason, discover meaning, generalize**, or **learn from past experience**

https://www.britannica.com/technology/artificial-intelligence

Can we trust Artificial Intelligence?

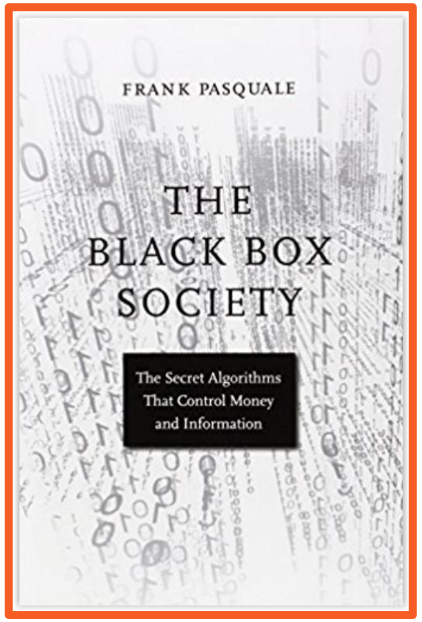
Angela Merkel
(German Chancellor, G8 President in 2007)

Referring to top credit rating agencies...

- *In the future it should be clear what the basis of their ratings of companies is.* ***There can't be some black-box from which something comes out and which no one understands****.*
  (Reuters, August 2007)

- **The Black Box Society: The Secret Algorithms that Control Money and Information**
  - F. Pasquale (2015), Cambridge, MA: Harvard University Press

"An **intelligible society** would assure that key decisions of its most important firms are **fair**, **nondiscriminatory**, and **open to criticism**. **Silicon Valley and Wall Street need to accept as much accountability as they impose on others**"



Can we trust Artificial Intelligence?

Sent to Prison by a Software Program's Secret Algorithms

Chief Justice John G. Roberts Jr., center, recently said that the day of using artificial intelligence in courtrooms was already here, "and it's putting a significant strain on how the judiciary goes about doing things." Stephen Crowley/The New York Times

https://www.youtube.com/watch?v=Gi4YeRqfb24

http://www.equivant.com/solutions/inmate-classification

https://www.nytimes.com/2017/05/01/us/politics/sent-to-prison-by-a-software-programs-secret-algorithms.html

Can we trust Artificial Intelligence?

https://www.youtube.com/watch?v=Onm6Sb3Pb2Y

Black Mirror in Beijing … China's new social credit scoring system is being implemented right now

March 4, 2019

China plans to rank all its citizens based on their "social credit" by 2020. Citizens will be rewarded or punished according to their scores. Like private financial credit scores, a person's social score can move up and down according to their behaviour.

The program is due to be fully operational nationwide by 2020, but is being piloted for millions of people across the country already. By the end of next year the scheme will be mandatory.

https://www.thegeniusworks.com/2019/03/chinas-black-mirror-reality-the-social-credit-scoring-is-being-implemented-rapidly/

Jose M. Alonso
https://citius.usc.es/v/jose-maria-alonso-moral

What is Explainable Artificial Intelligence?

Explainable Models
+ Interpretable Models
+ Hybrid ML-KB Models
+ Model agnostic Explainers
+ Counterfactual Explanations

Explanation Interfaces
+ HCI
+ Multimodal Communication Strategies
+ Behavior Change Support Systems

Psychology and Evaluation of Explanation
+ Expert Systems
+ Cognitive Computing
+ Argumentation Theory
+ Discourse History
+ Human Evaluation

explanation

The system takes input from the current task and makes a recommendation, decision, or action

The system provides an explanation to the user that justifies its recommendation, decision, or action

MARIE CURIE ACTIONS

NL4XAI

NoBIAS

erc European Research Council

GDPR LAW

AI4EU Advancing Europe through collaboration in AI

20 million euros
21 countries
70 partners
https://www.ai4eu.eu/

HUMANE AI

FET Flagship Project Proposal
Ethical and Trustworthy AI
17 countries
35 partners
http://www.humane-ai.eu/

CLAIRE
https://claire-ai.org/

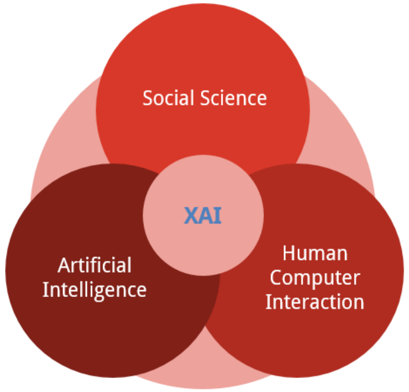
DARPA Challenge on eXplainable Artificial Intelligence (XAI) (August 2016, DARPA-BAA-16-53)
http://www.darpa.mil/program/explainable-artificial-intelligence

D. Gunning, D. Aha, "**DARPA's Explainable Artificial Intelligence (XAI) Program**", AI Magazine, 40(2):44-58, 2019, https://doi.org/10.1609/aimag.v40i2.2850

Who/What is XAI for?

Who?

- ✓ Regular (lay) user
- ✓ Expert user
- ✓ Developers
- ✓ External entity

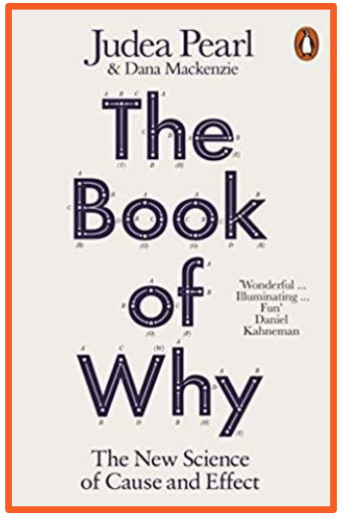


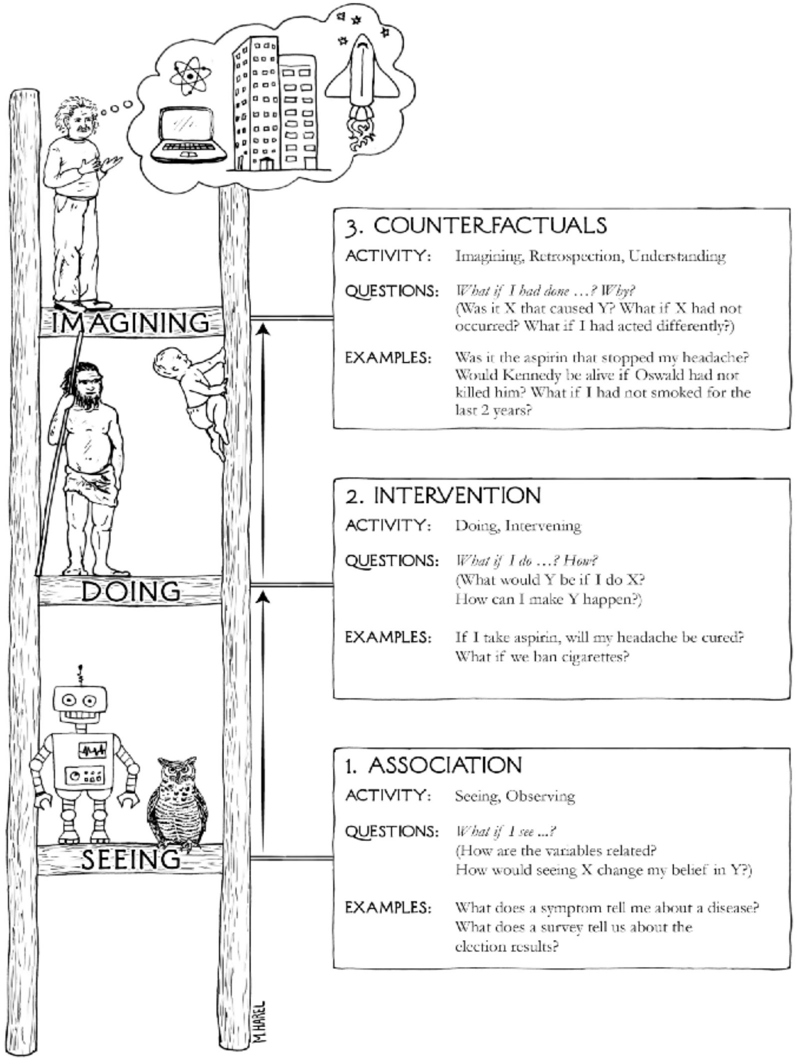
What?

- ❖ To **justify system decisions** (so that humans can accept them)
- ❖ To **explain system decisions** (to guarantee safety concerns are met)
- ❖ To **build trust** in system decisions (especially if a mistake is suspected or the human operator does not have experience with the system)
- ❖ To **explain system's choices** (to ensure fair, ethical, and/or legal decisions are made)
- ❖ To **explain the system's choices** (to better evaluate or debug the system in previously unconsidered situations)
- ❖ To **facilitate Knowledge / scientific discovery**

T. Miller, "**Explanation in Artificial Intelligence: Insights from the Social Sciences**", Artificial Intelligence, 267:1-38, 2019, https://doi.org/10.1016/j.artint.2018.07.007

Explainability – WHY Questions ?





Explanation is an answer to a "why?" question

- "Why?" = "How come?"
  - Why planets are spherical?
- "Why?" = "What for?"
  - Why ball bearings are spherical?

Most often than not, "why?" questions are contrastive

- Why did Elizabeth open the door? (rather than leave it closed)

Interpretability & Explainability

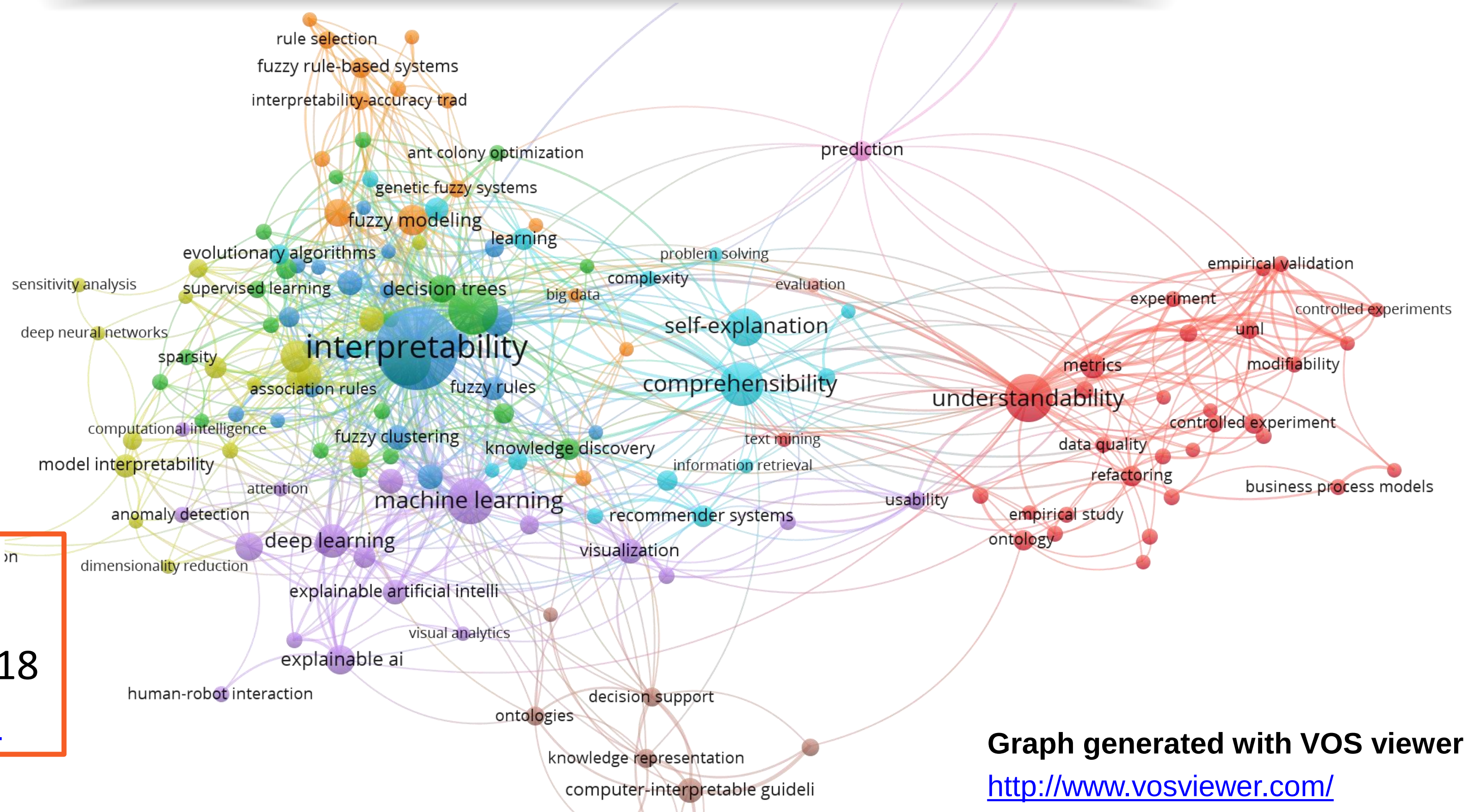
```
Q1 = TITLE (``*interpretab*'') OR TITLE (``*comprehensib*'')
OR TITLE (``*understandab*'') OR TITLE (``*explainab*'')
OR TITLE (``*self-explanat*'') OR KEY (``*interpretab*'')
OR KEY (``*comprehensib*'') OR KEY (``*understandab*'')
OR KEY (``*explainab*'') OR KEY (``*self-explanat*'')
```

❑ **Adv. Search query:**

Analysis restricted to
- 2000-2020

and to the subject areas of
- Computer Science
- Mathematics
- Engineering

6732 publications considered



[UPDATED]

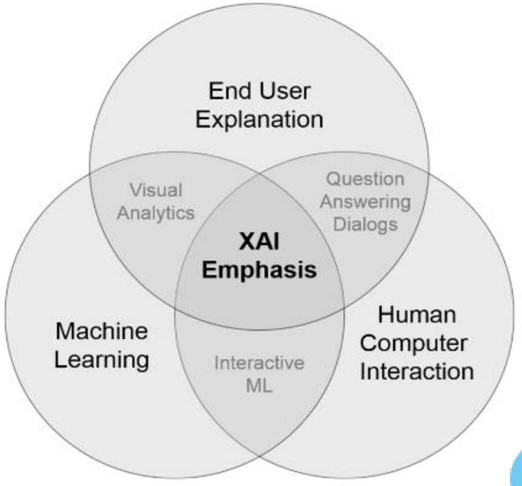
J. M. Alonso, C. Castiello, C. Mencar,
"**A Bibliometric Analysis of the Explainable Artificial Intelligence Research Field**", IPMU2018
https://doi.org/10.1007/978-3-319-91473-2_1

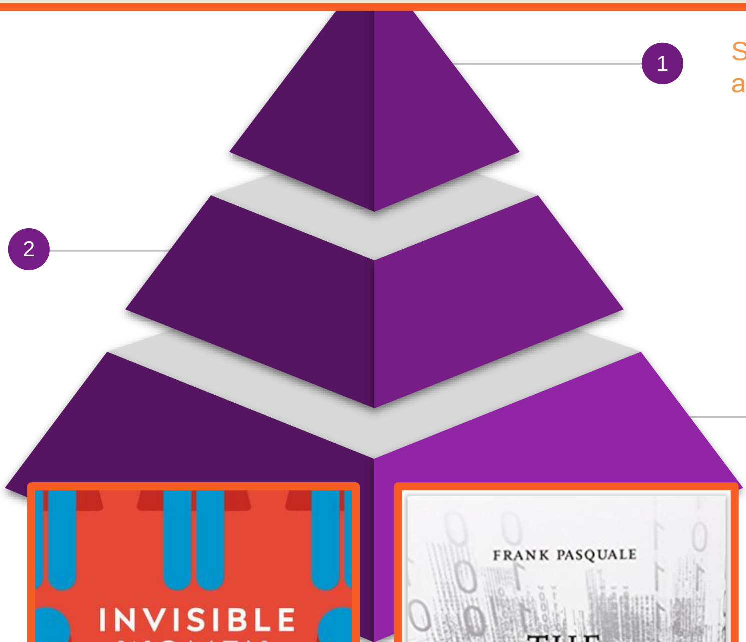
Graph generated with VOS viewer
http://www.vosviewer.com/

Centro Singular de Investigación en Tecnoloxías Intelixentes

Interpretability & Explainability

Transparency = Say what you do !
Coherence = Do what you say !





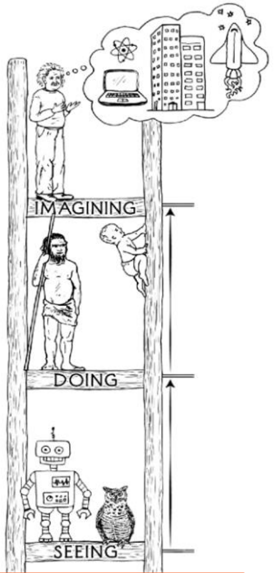
1 Systems capable of answering "why?" questions

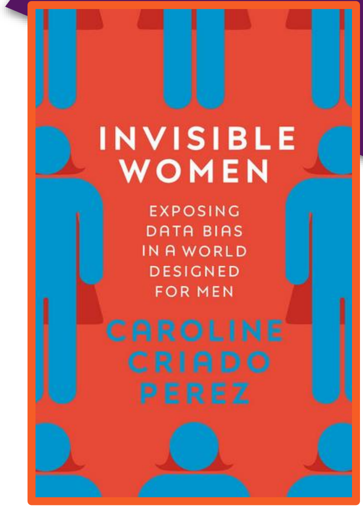
Interpretable systems

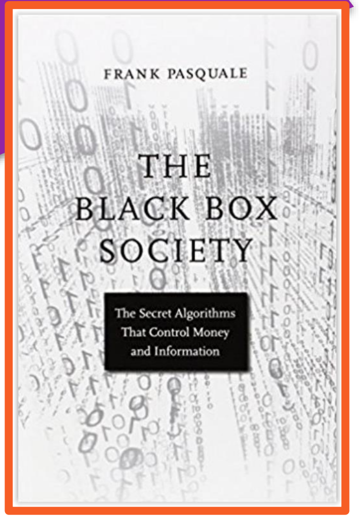
2 Systems whose structure and behavior can be understood

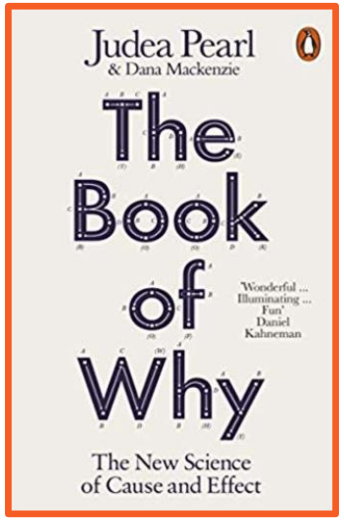
Machine learning

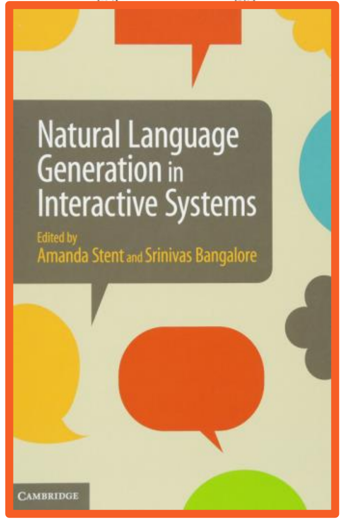
3 Systems that self-adapt to data





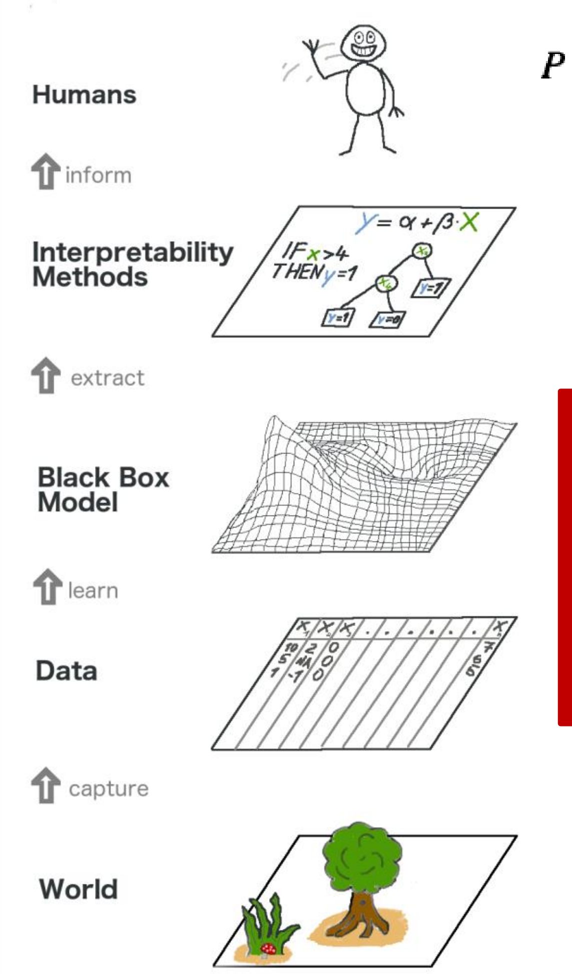






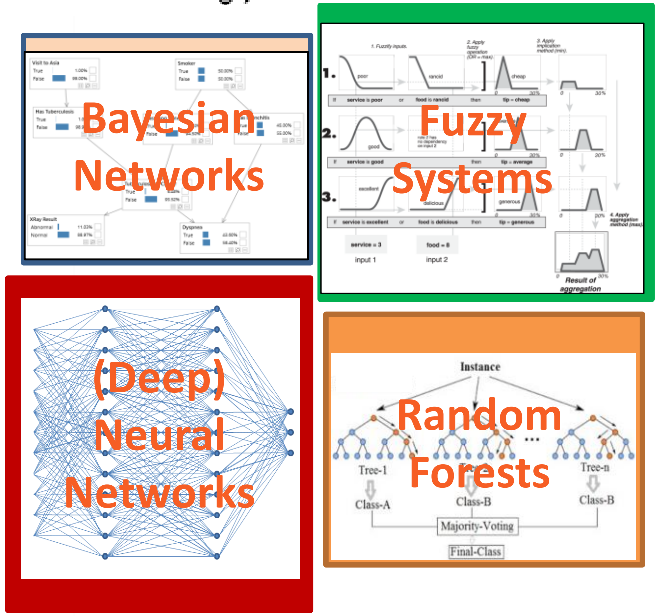
Jose M. Alonso
https://citius.usc.es/v/jose-maria-alonso-moral

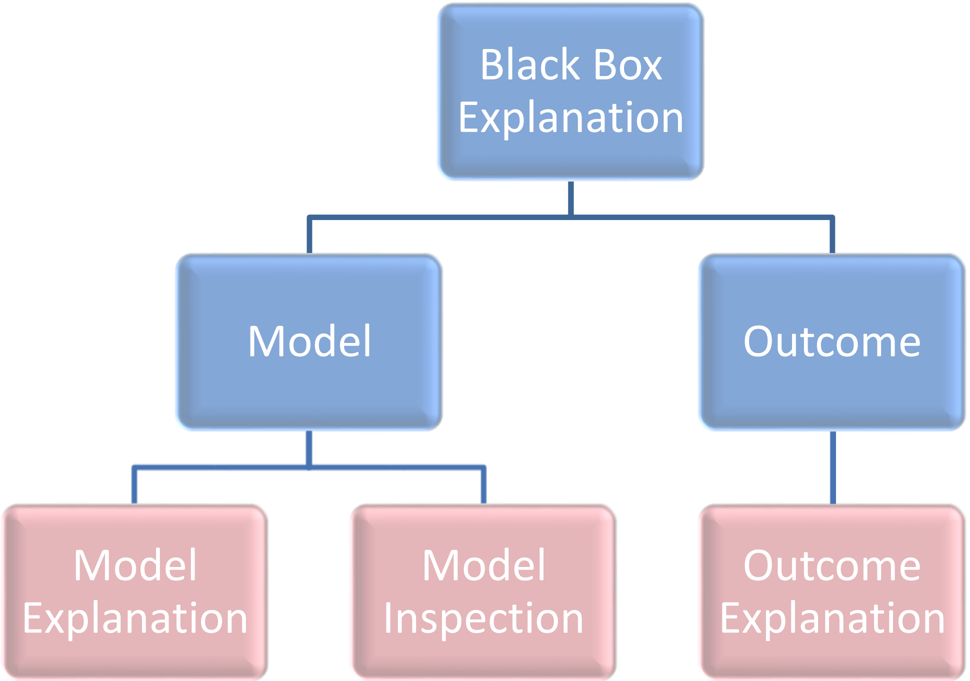
Interpretability & Explainability



$$P(x \mid y) = \frac{P(x, y)}{P(y)}$$

IF-THEN





R. Guidotti, et al, "**A Survey of Methods for Explaining Black Box Models**", ACM Computing Surveys, 51(5), 93:1-93:42, 2018, https://doi.org/10.1145/3236009

C. Rudin, "**Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead**", Nature Machine Intelligence 1(5):206–215, 2019, https://dx.doi.org/10.1038/s42256-019-0048-x

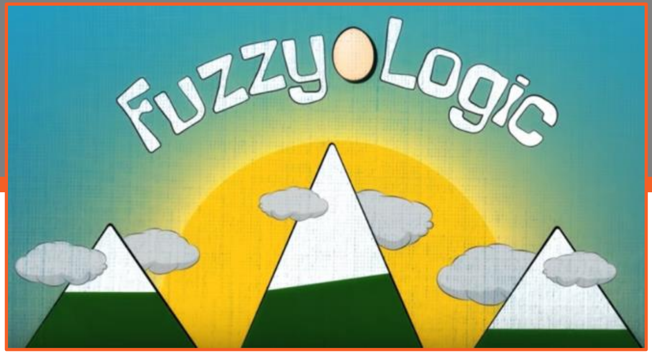
(*) Left Picture taken from https://christophm.github.io/interpretable-ml-book/agnostic.html

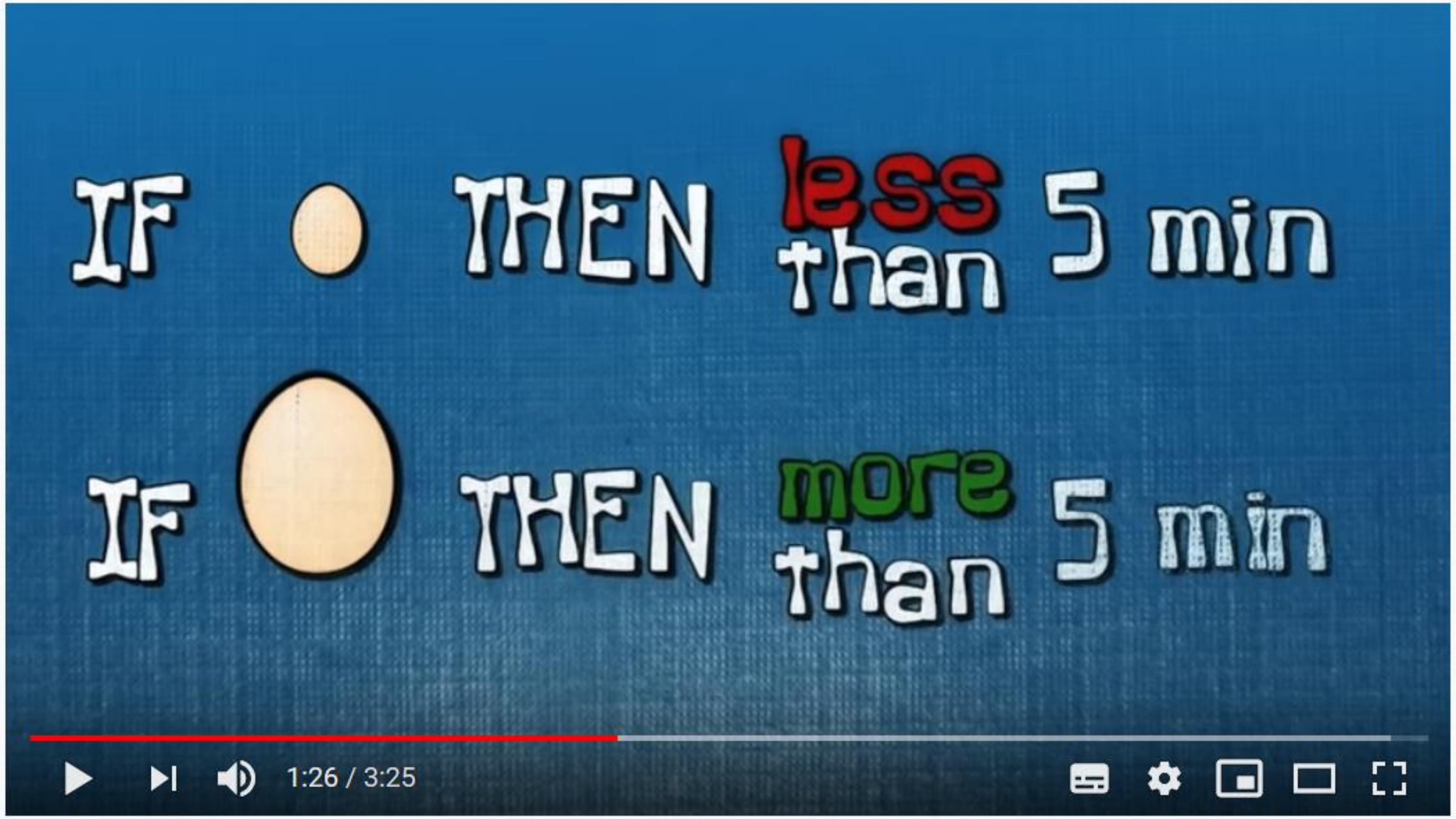
Interpretable Fuzzy Systems

Jose M. Alonso
https://citius.usc.es/v/jose-maria-alonso-moral

Fuzzy Sets and Systems

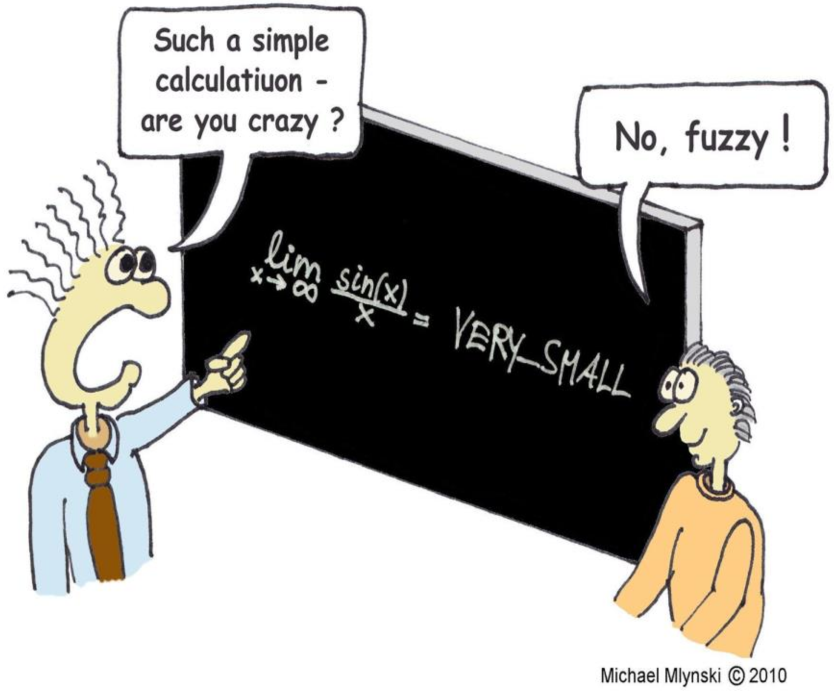
A matter of degree – handling uncertainty and truth values



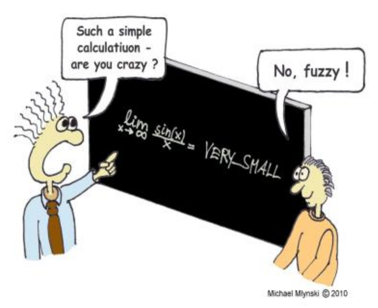


An Egg-Boiling Fuzzy Logic Robot

http://www.youtube.com/watch?v=J_Q5X0nTmrA



Historical Overview on Linguistic and Approximative Fuzzy Modeling



http://www.youtube.com/watch?v=2ScTwFCcXGo

[1965] **Fuzzy Sets**

[1965 – 1990] **Interpretability (I)** - LFM

- ❑ Fuzzy Reasoning (dealing with uncertainty)
- ❑ Simple linguistic variables and rules (high interpretability)
- ❑ Expert knowledge (Fuzzy Control and Expert Systems)

[1990 – 2000] **Accuracy (A)** - AFM

- ❑ Complex fuzzy rules with high accuracy
- ❑ Induced knowledge (Machine Learning, Hybrid Systems)

[2000 – 2014] **I-A Trade-off** (LFM + AFM)

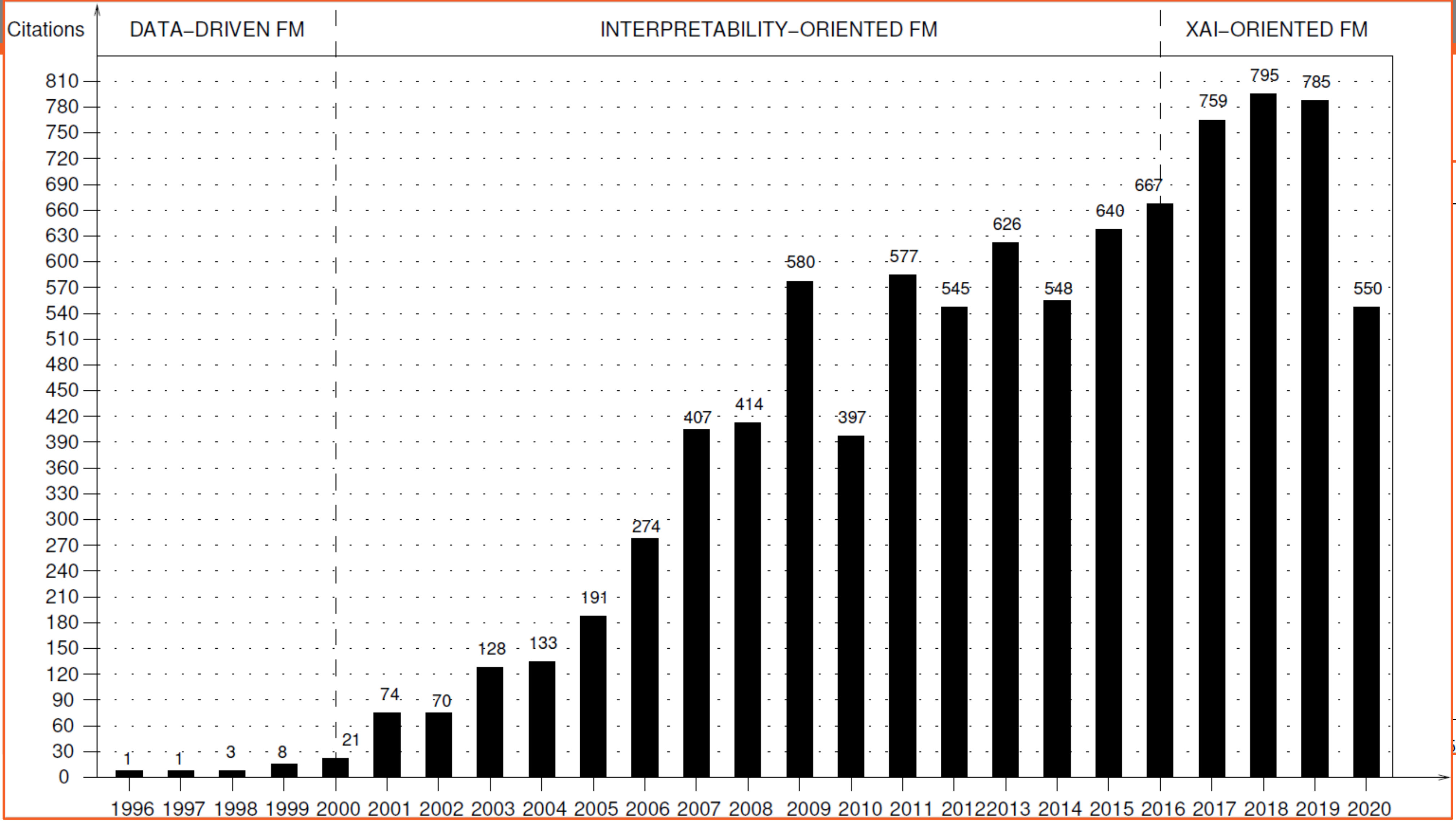
- ❑ Simple linguistic rules with high accuracy
- ❑ Expert + Induced knowledge, Multi-objective design

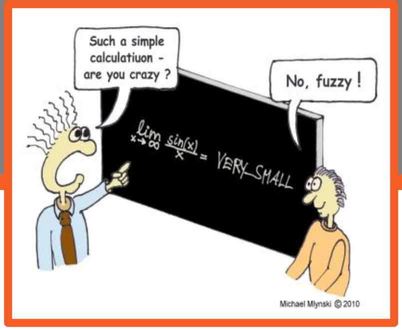
[2014 – 2016] Internet of Things, Big Data, Social Networks, Industry 4.0

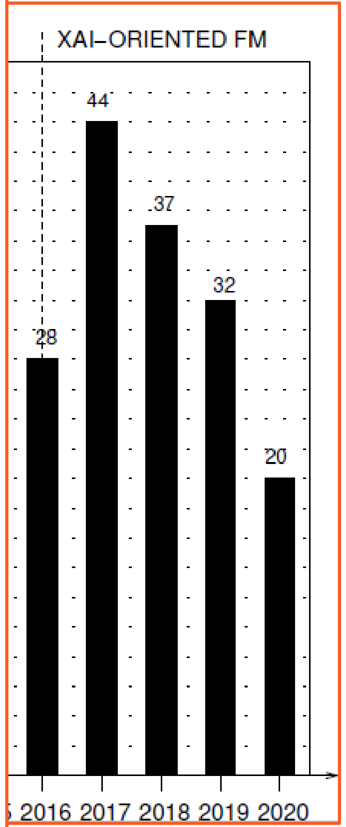
[2017 – 2020] **Explainable Artificial Intelligence**

- **Linguistic Summarization of Data** (Yager 1990)
- **Computational Theory of Perceptions** (Zadeh 2001)
  "From Computing with Numbers to Computing with Words", "From Manipulation of Measurements to Manipulations of Perceptions"

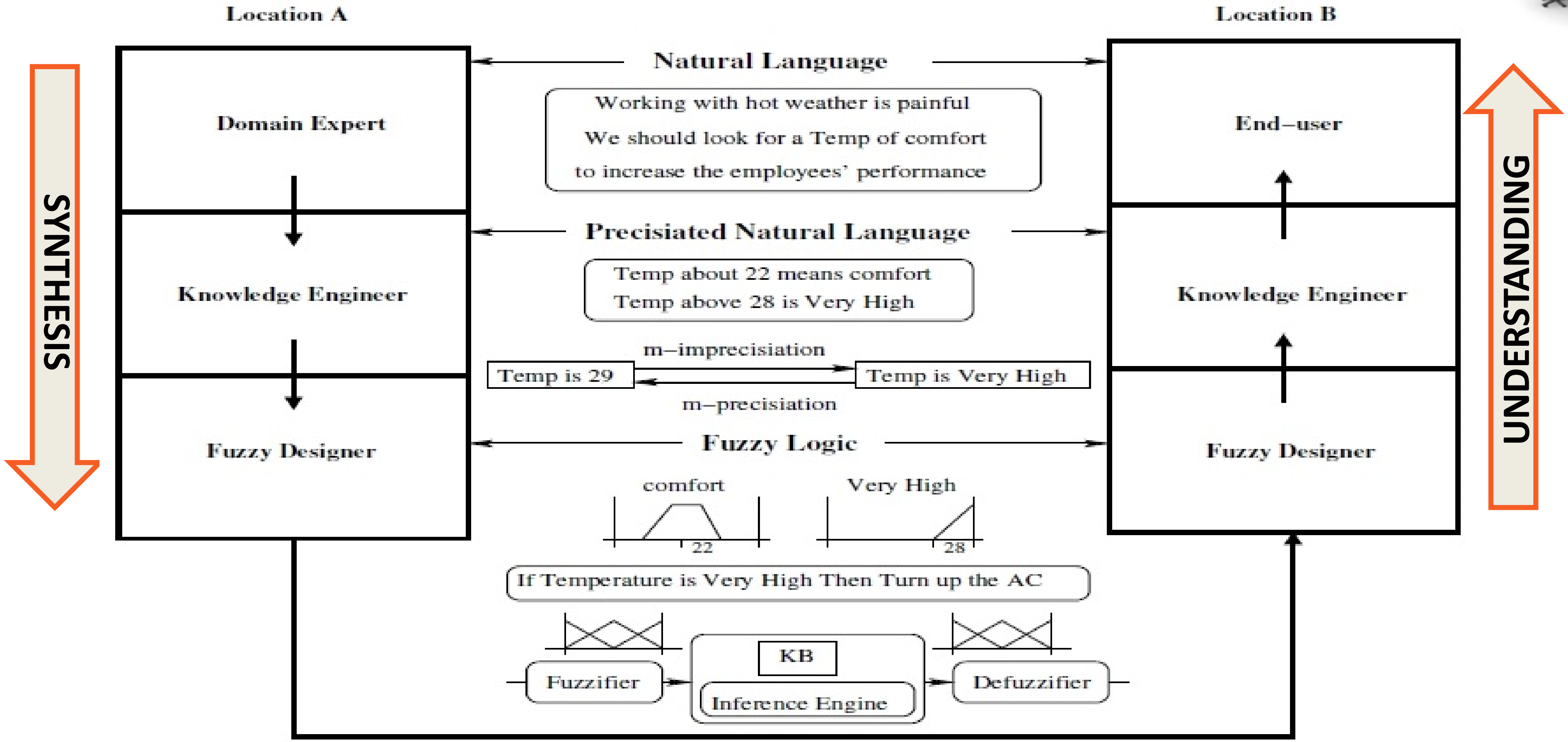
Bibliometric Overview on Interpretable Fuzzy Systems





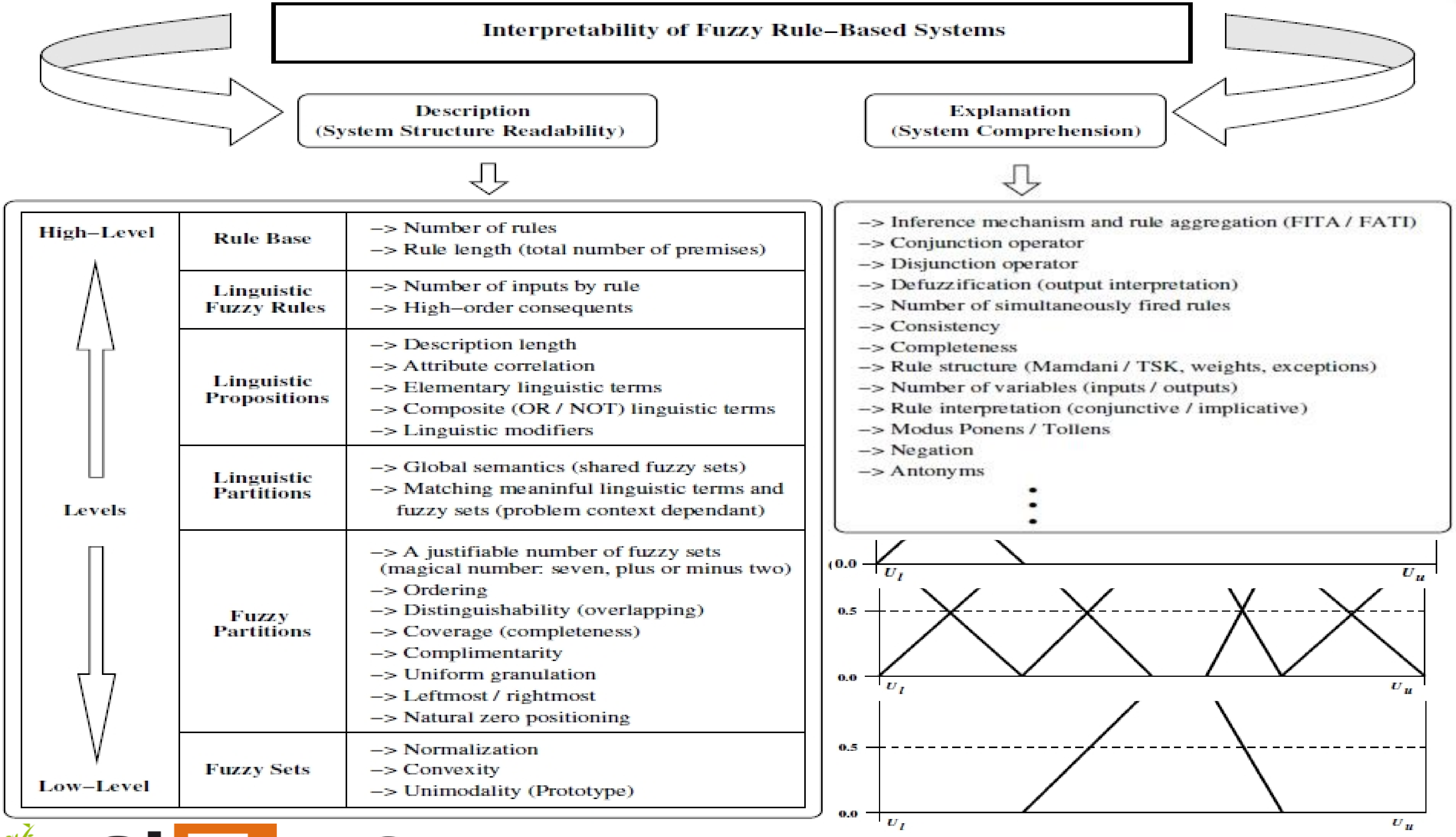


Building Interpretable Fuzzy Systems
A matter of careful human-centered design



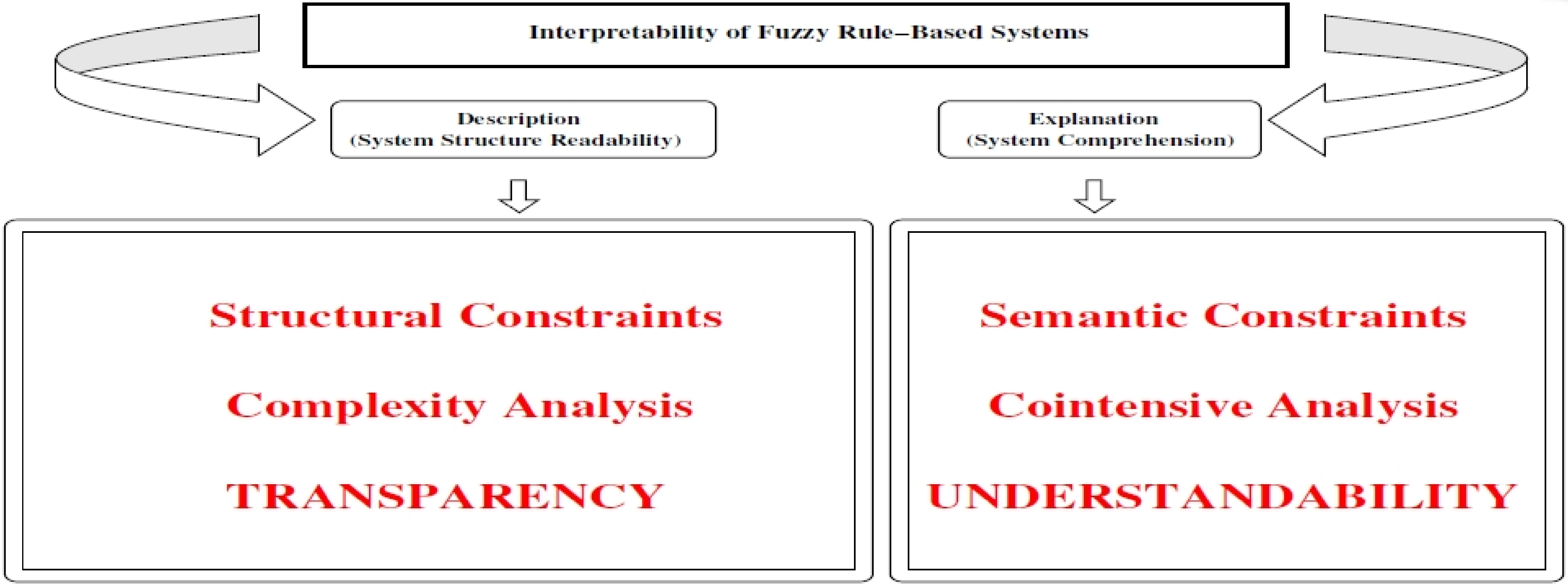
Jose M. Alonso, C. Castiello, C. Mencar, "**Interpretability of fuzzy systems: Current research trends and prospects**", Springer Handbook of Computational Intelligence, pp. 219-237, 2015, https://dx.doi.org/10.1007/978-3-662-43505-2_14

Building Interpretable Fuzzy Systems
A matter of careful human-centered design



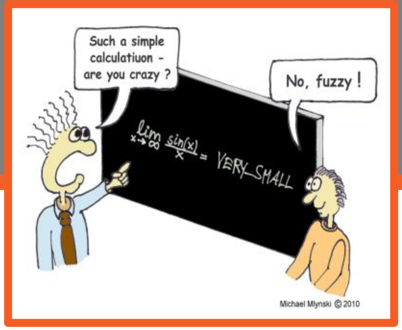
Building Interpretable Fuzzy Systems

A matter of careful human-centered design



The purpose of building descriptions in Natural Language is to provide end-users with textual information which is expected to be easy to read and to understand

Interpretability and Language



- **Interpretability is essential for effective communication**
- **How to organize a message** (oral speech / text written) **to become interpretable?**

  Thinking on the expected audience's background (**Comunicative Goal** + **User Model**)

  And keeping in mind:

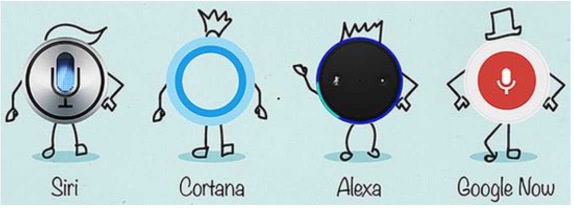
  - **Paul Grice's Maxims** (Logic and conversation, 1975): Quality, Quantity, Relation (relevance), and Manner (brief, orderly)
  - **Occam's Razor Principle** (14th-century): Assuming two explanations are equivalent in informative terms then the simplest one is the best
  - **Inquiries into Truth and Interpretation** (Oxford 1985)
  - **Meaning Holism and Interpretability** (The Philosophical Quaterly 1991): *"... an interpreter who finds a speaker mistaken in one case might be obliged by meaning holism to find him mistaken in most cases... the possibility of massive error threatens interpretability... there can be no language that is uninterpretable..."*
  - **Minimum Description Length Principle** (Zemel, 1998)

Comprehensibility Postulate (R.S. Michalski, 1983)

*"The results of **computer induction** should be **symbolic descriptions** of given entities, **semantically and structurally similar** to those a **human expert** might produce observing the same entities"*

Natural Language Technology

- **Natural Language Processing** (NLP):
  - Natural Language **Understanding** (NLU): **analytics from texts**
  - Natural Language **Generation** (NLG): **texts from other data sources**
- As a research field, **NLG** is in development for **more than 25 years**
  - **Text-to-Text (T2T)**
  - **Dialogue Systems**
  - **Computational Creativity**
  - **Data-to-Text (D2T)**



NATURAL LANGUAGE PROCESSING

R. Dale, "**The commercial NLP landscape in 2017**", Natural Language in Engineering, 23(4):641-647, 2017, http://dx.doi.org/10.1017/S1351324917000237

A. Gatt, E. Krahmer, "**Survey of the State of the Art in Natural Language Generation: Core Tasks, Applications and Evaluation**", Journal of Artificial Intelligence Research, 61:65–170, 2018, https://doi.org/10.1613/jair.5477

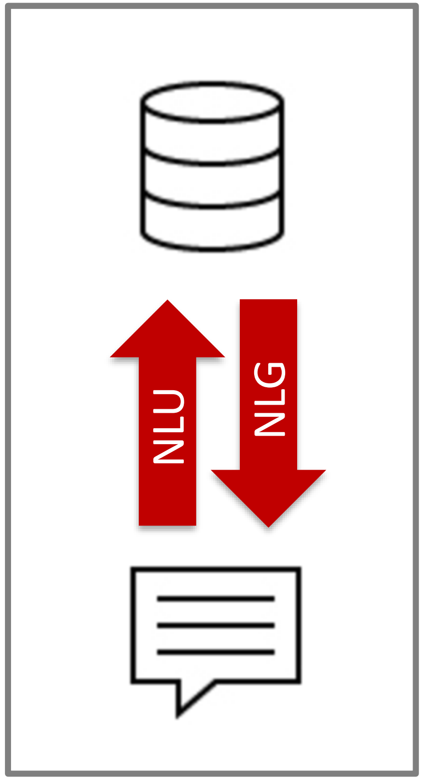
Jose M. Alonso
https://citius.usc.es/v/jose-maria-alonso-moral

Natural Language Technology

Natural Language Generation (NLG)

- **Text-to-Text (T2T)**
  - Generation of coherent texts from other texts (includes NLU)
- **Dialogue Systems**
  - Dialog generated from texts provided by users or bots (includes NLU)
- **Computational Creativity**
  - Generation of histories, tales, poems, etc.
- **Data-to-Text (D2T)**
  - Text generation from numerical or symbolic data sets or series

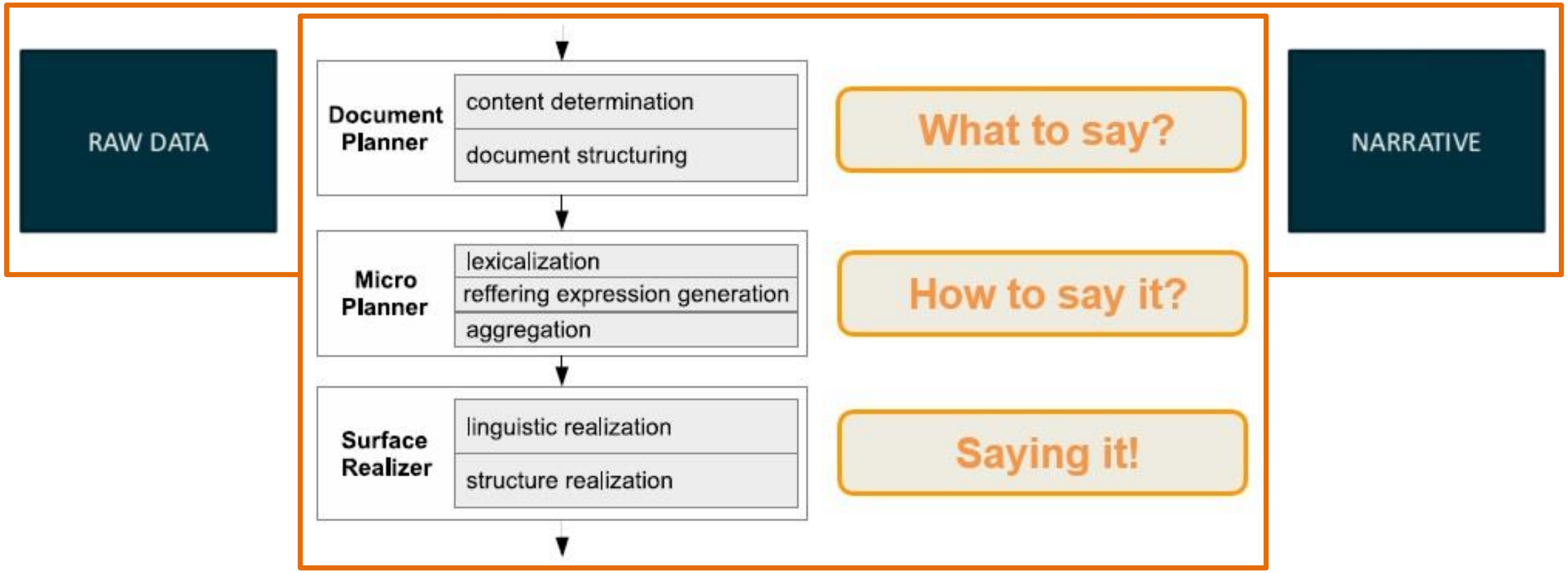
NATURAL LANGUAGE PROCESSING



A. Gatt, E. Krahmer, "**Survey of the State of the Art in Natural Language Generation: Core Tasks, Applications and Evaluation**", Journal of Artificial Intelligence Research, 61:65–170, 2018, https://doi.org/10.1613/jair.5477

Data-to-text (D2T) natural language generation systems are able to process (huge) quantities of data and convert them into comprehensible texts, which contain relevant information for human users.

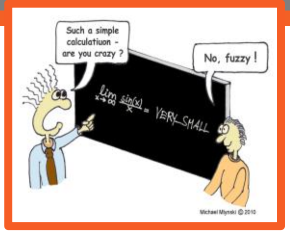
Ehud Reiter



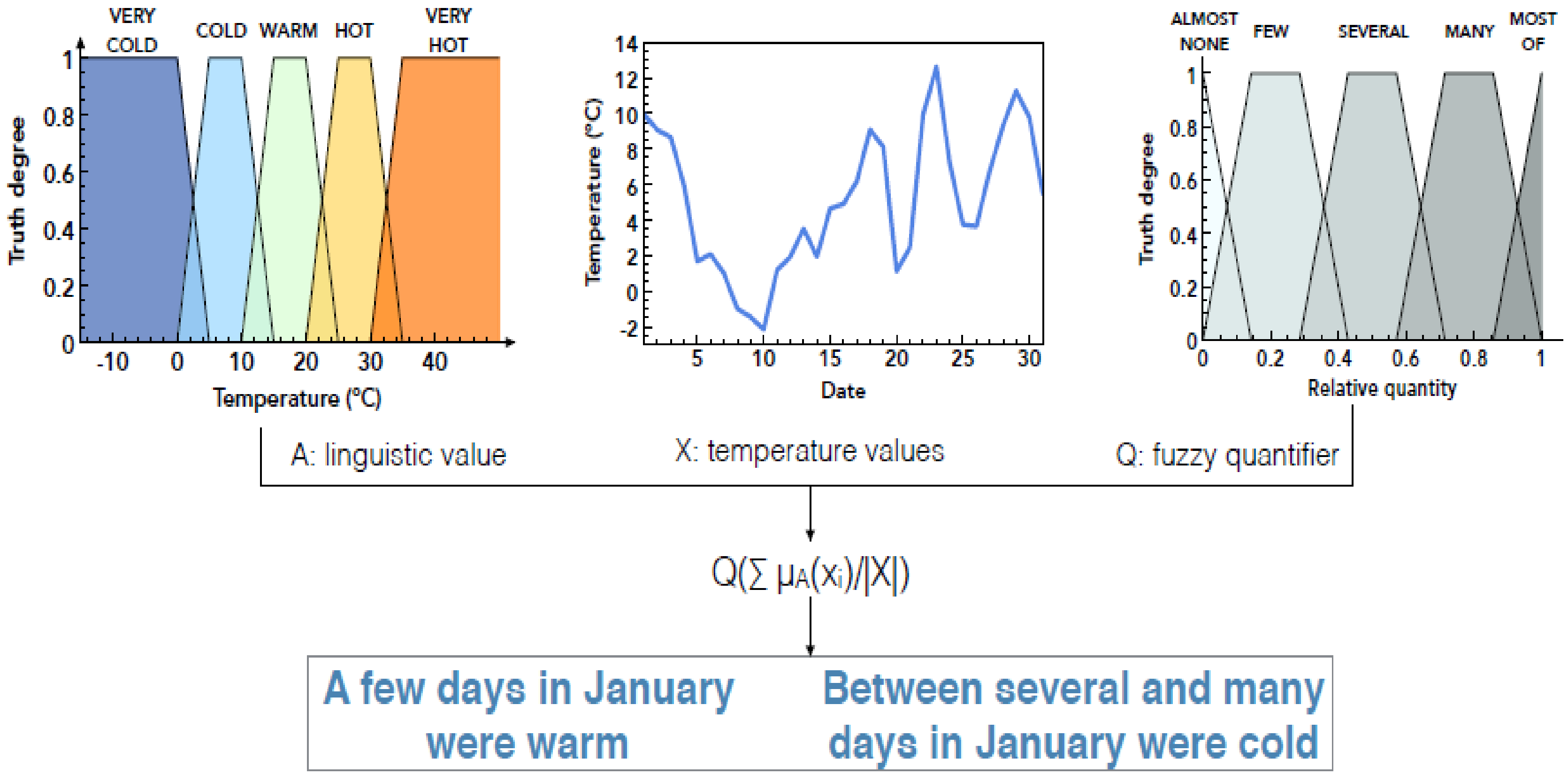
E. Reiter and R. Dale, "**Building Natural-Language Generation Systems**", Cambridge University Press, 2000

NLG & Fuzzy Logic

From numbers to sentences managing human vagueness



- **Linguistic summarization** (Yager, 1982)
- **Zadeh's protoforms, fuzzy quantified sentences** (Zadeh, 1983)

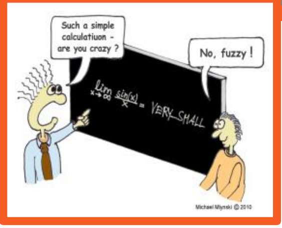


$$card_E(A) = \frac{\sum_{e \in E} \mu_A(e)}{|E|}$$

$$card_E(A) = \sum_{e \in E} \mu_A(e)$$

NLG & Fuzzy Logic

From numbers to sentences managing human vagueness



- **Protoforms are not (in general) texts ready to be conveyed for human consumption (except for non-trivial cases)**

Content

Q Xs are A
Most of coming days are cloudy
Most of the coming days will be cloudy
The skies will be predominantly cloudy for the coming short-term period

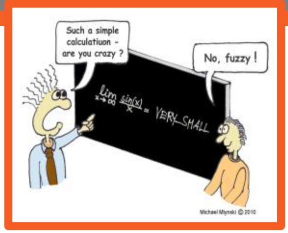
Language

A. Ramos-Soto, A. Bugarín, S. Barro, “**Fuzzy sets across the natural language generation pipeline**”, Progress in Artificial Intelligence, 5(4):261-276, 2016, https://doi.org/10.1007/s13748-016-0097-x

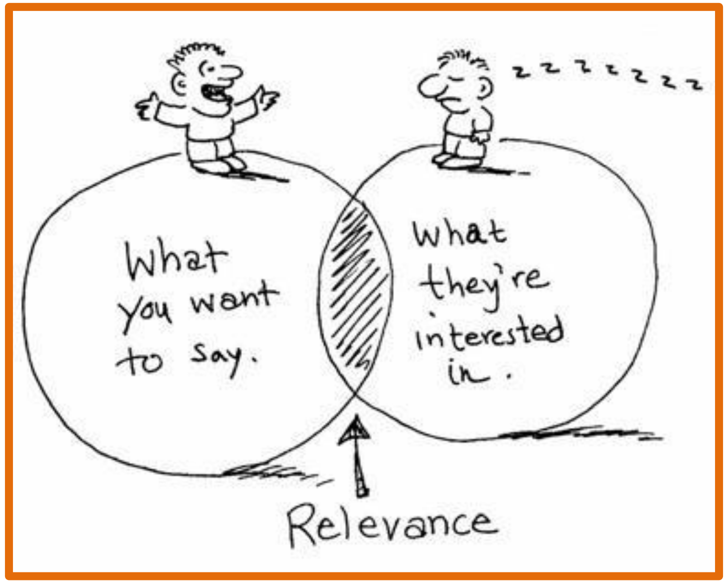
A. Ramos-Soto, A. Bugarín, S. Barro, “**On the role of linguistic descriptions of data in the building of natural language generation systems**”, Fuzzy Sets and Systems, 285, 31-51, 2016, https://doi.org/10.1016/j.fss.2015.06.019

A. Ramos-Soto, Jose M. Alonso, E. Reiter, K. van Deemter, A. Gatt, “**Fuzzy-Based Language Grounding of Geographical References: From Writers to Readers**”, International Journal of Computational Intelligence Systems, 12(2):970-983, 2019, https://dx.doi.org/10.2991/ijcis.d.190826.002

NLG & Fuzzy Logic
Computing with Words and Perceptions



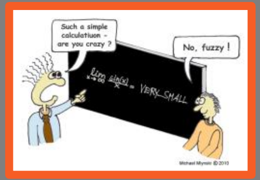
- **Computational Theory of Perceptions**
  - "From computing with numbers to computing with words" (Zadeh, 1999)
  - "Toward a perception-based theory of probabilistic reasoning" (Zadeh, 2002)
- **Computing with Words and NLG** (Kaczprycz & Zadrozny, 2010)
- **Linguistic Description of Complex Phenomena (LDCP)** (Trivino & Sugeno, 2013)
- **LDCP: Applications with Big Data** (Conde-Clemente, 2017)

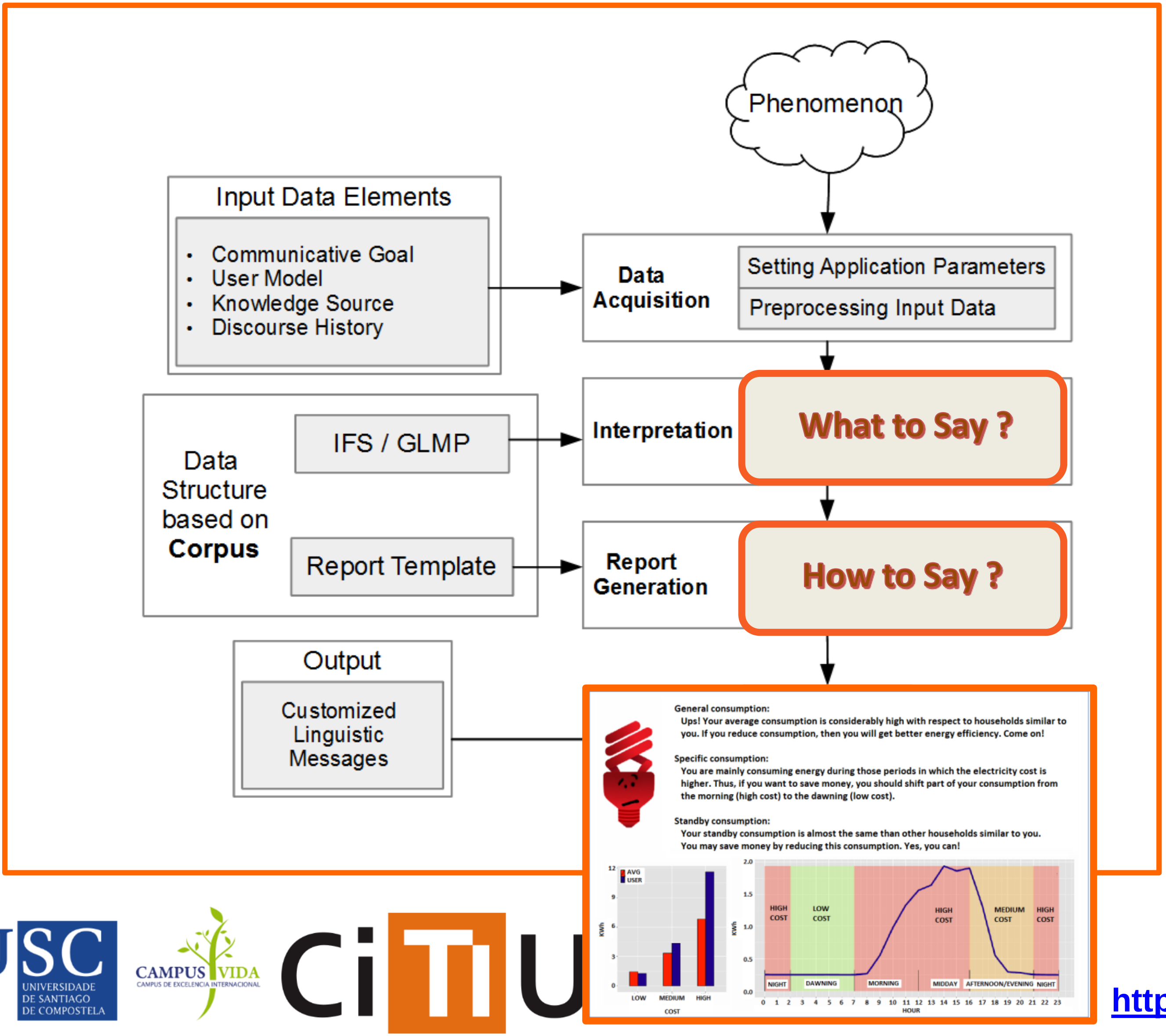


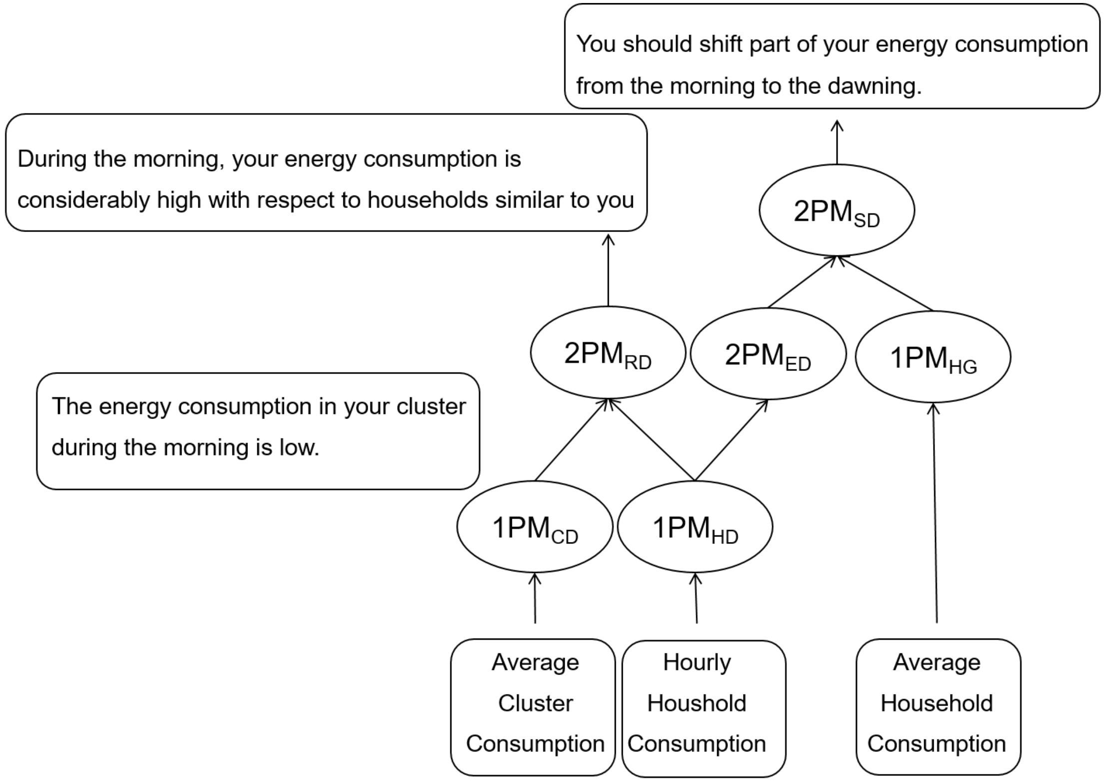
P. Conde-Clemente, G. Trivino, Jose M. Alonso, "**Generating Automatic Linguistic Descriptions with Big Data**", Information Sciences, 380:12-30, 2017, https://dx.doi.org/10.1016/j.ins.2016.11.002

NLG & Fuzzy Logic

Linguistic Description of Complex Phenomena (LDCP)







P. Conde-Clemente, Jose M. Alonso, G. Trivino, "**Towards automatic generation of linguistic advice for saving energy at home**", Soft Computing, 1-15, 2016
https://doi.org/10.1007/s00500-016-2430-5
https://www.youtube.com/watch?v=TzTS388T_-U

NATCONSUMERS

Jose M. Alonso
https://citius.usc.es/v/jose-maria-alonso-moral

SCIS&ISIS2020

Explainable Fuzzy Systems

Centro Singular de Investigación en Tecnoloxías Intelixentes

Jose M. Alonso
https://citius.usc.es/v/jose-maria-alonso-moral

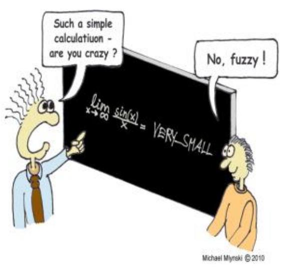
SCIS&ISIS2020

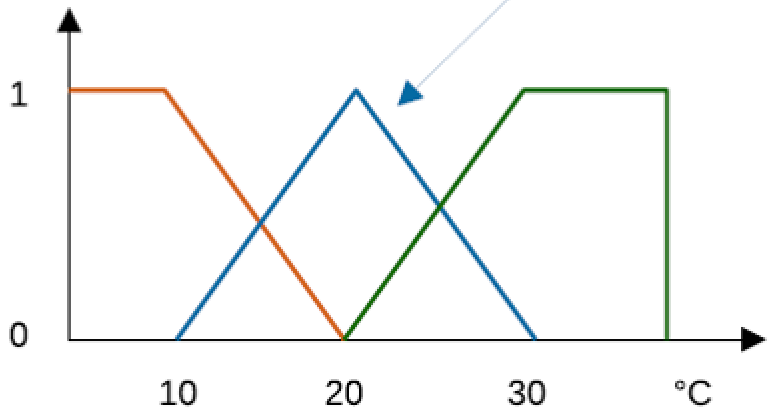
Explainable Fuzzy Systems (EXFS) = IFS + NLU + NLG + HCI

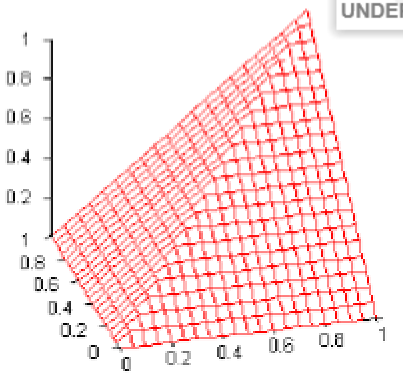
Ready to answer WHY questions ?

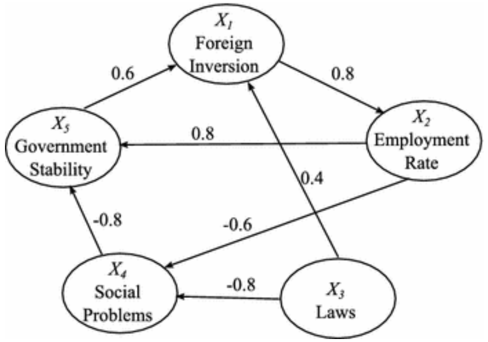
- Fuzzy-grounded Knowledge Representation and Reasoning
- Computing with Words
- Fuzzy Cognitive Maps
- CHC models
  - **Consequences**
  - **Hypothesis**
  - **Conjectures**
    - ✓ **Speculations**
    - ✓ **Refutations**

IF Temperature IS Warm AND ...









E. Trillas, I. García-Honrado, A. Pradera, "**Consequences and conjectures in preordered sets**", Information Sciences, 180(19):3573-3588, 2010, https://doi.org/10.1016/j.ins.2010.06.006

D. Dubois, H. Prade, "**What are fuzzy rules and how to use them**", Fuzzy Sets and Systems, 84(2):169–185, 1996, https://dx.doi.org/10.1016/0165-0114(96)00066-8

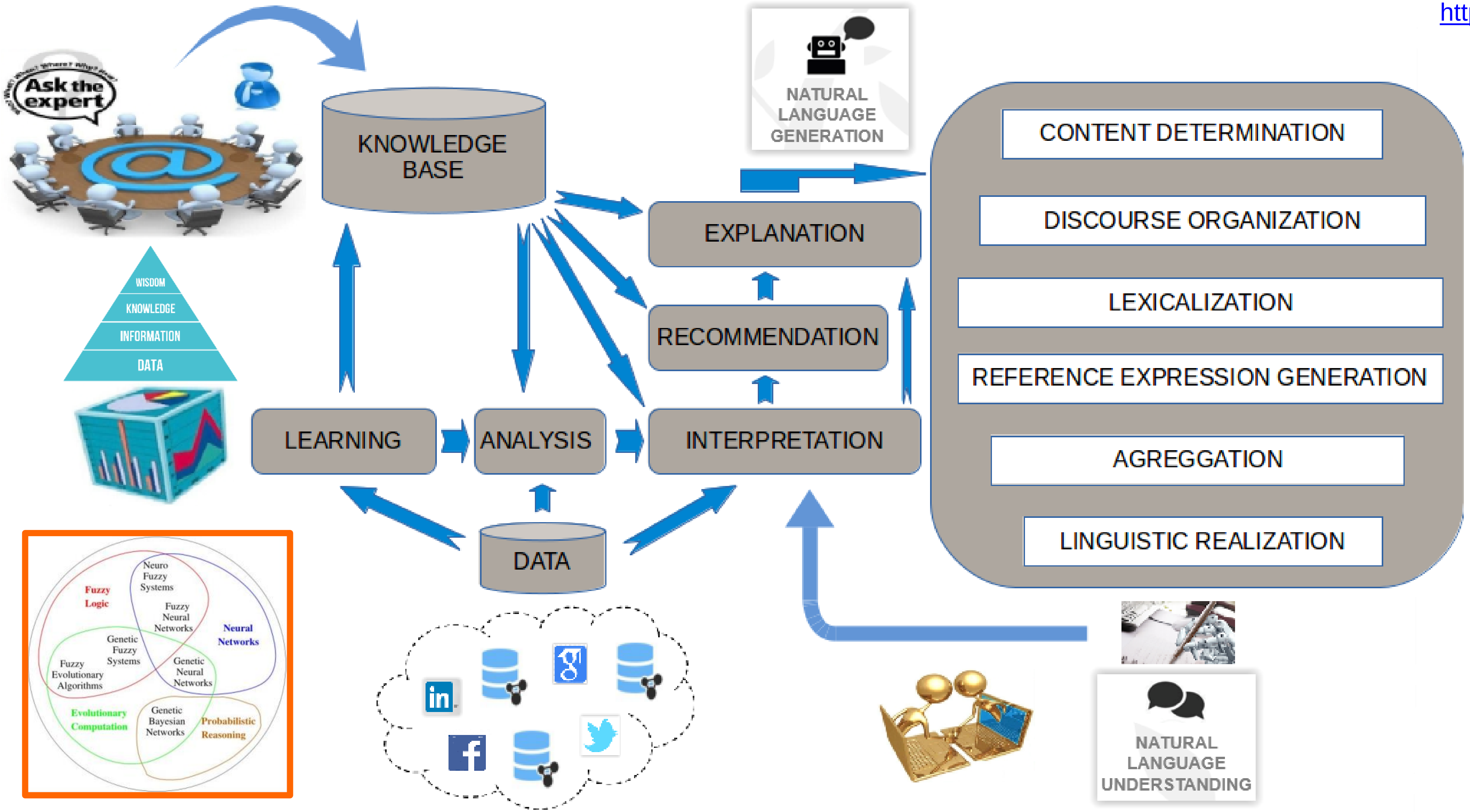
E. Trillas, L. Eciolaza, "**Fuzzy Logic: An Introductory Course for Engineering Students**", Springer, 2015, https://dx.doi.org/10.1007/978-3-319-14203-6

B. Kosko, "**Fuzzy Cognitive Maps**", International Journal of Man-Machine Studies, 24:65-75, 1986, https://doi.org/10.1016/S0020-7373(86)80040-2

How to build EXFS ?

rLDCP + GUAJE + JFML + ExpliClas

http://sci2s.ugr.es/es/fss



rLDCP

rLDCP package

R package for text generation from data

Home | Handbook | Downloads

Handbook

This section provides you with instructions for installing the rLDCP package and running the example "ComfortableRoom" which is described in the paper.

Here you can download a file with all instructions given in the section.

Installing the rLDCP package

The latest version of the package is in the Downloads tab. The binaries are currently only available for Windows. You should compile the source package for other operating systems.

First of all, the user has to download the package, and then its installation is as follows (Please, note that working with absolute path is required):

```
> install.packages('C:\\Downloads\\rLDCP_1.0.2.zip',repos=NULL)
package 'rLDCP' successfully unpacked and MD5 sums checked
```

Also, the user can install the package with RStudio. In the RStudio menu the user must select **Tools** and then **Install Packages...**. As result, the next popup window is open. The user must: (1) select that the package is installed from **Package Archive File (.zip; .tar.gz)**; (2) select the package path **C:/Downloads/rLDCP_1.0.2.zip**; and (3) give a directory to install the library or just go on with the default installation directory.

P. Conde-Clemente, Jose M. Alonso, G. Trivino, "**rLDCP: R package for text generation from data**", IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Naples, Italy, 2017, https://dx.doi.org/10.1109/FUZZ-IEEE.2017.8015487
http://www.phedes.com/rLDCP/

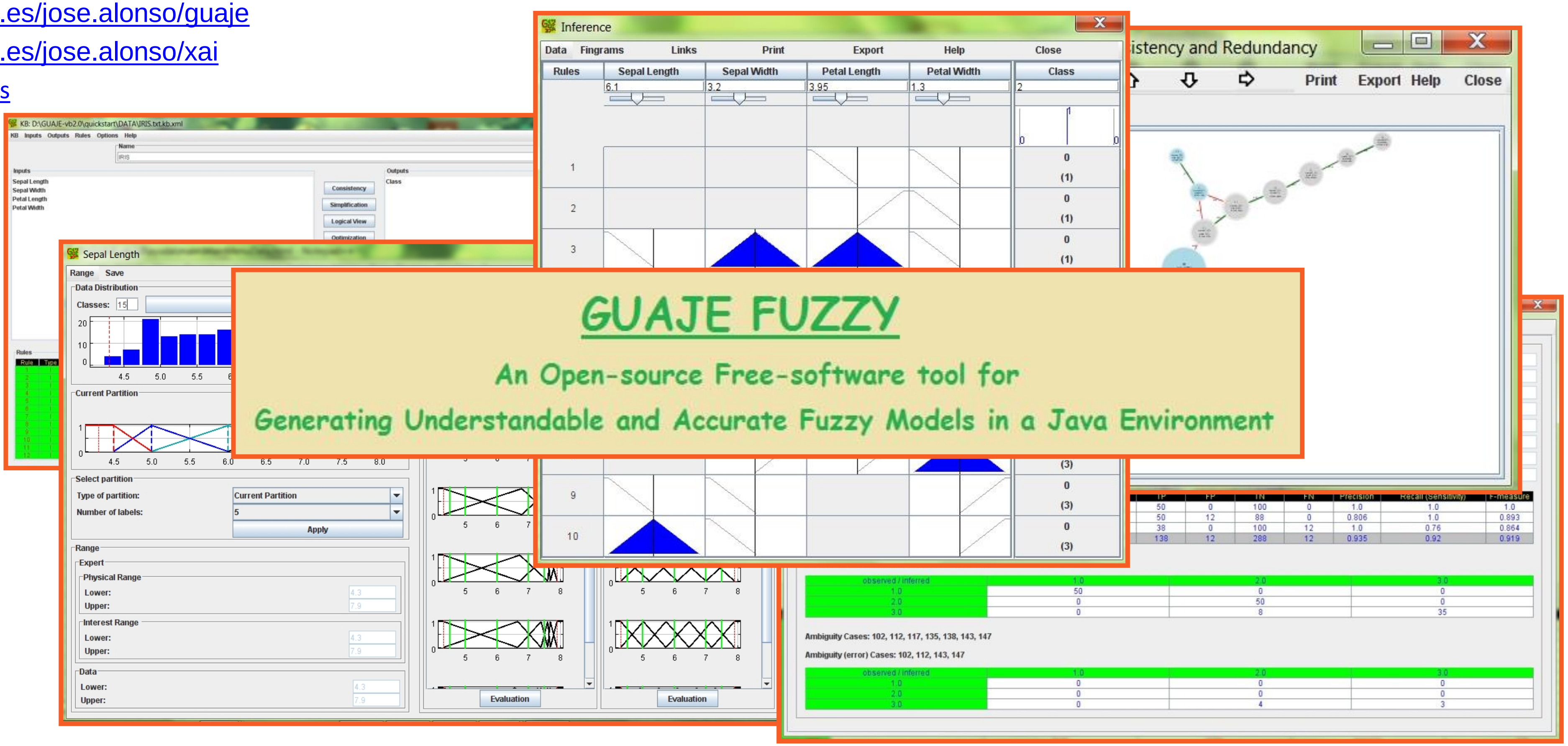
GUAJE

Expert + Data: Fuzzy Rule-based Systems with Global Semantics

https://gitlab.citius.usc.es/jose.alonso/guaje

https://gitlab.citius.usc.es/jose.alonso/xai

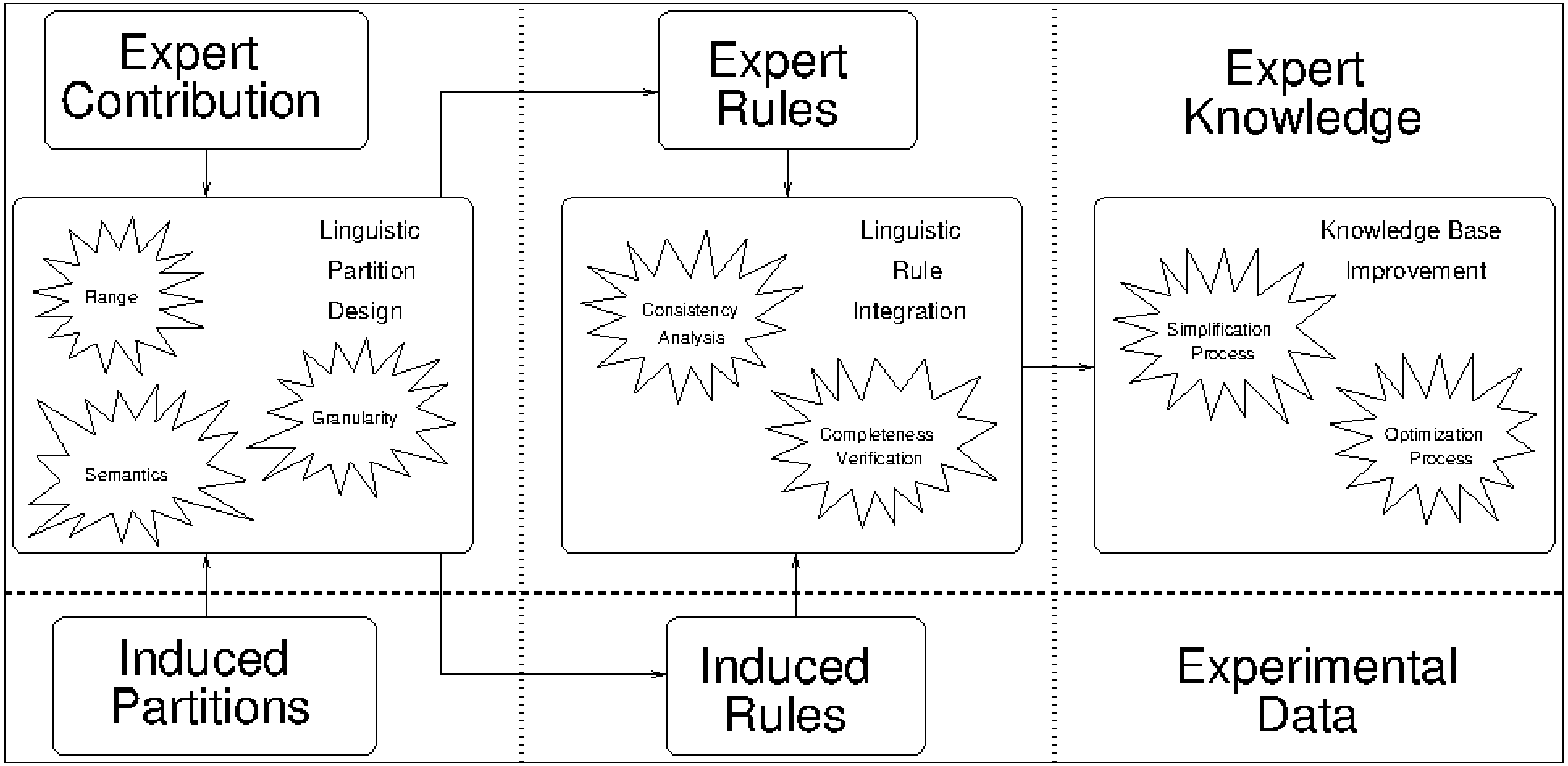
http://sci2s.ugr.es/es/fss

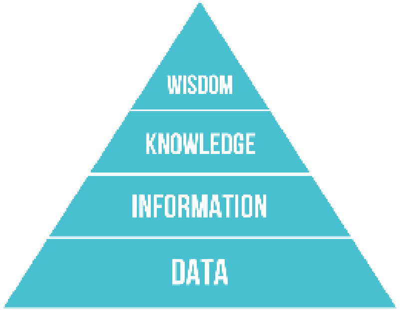


Jose M. Alonso, L. Magdalena, "**Generating Understandable and Accurate Fuzzy Rule-Based Systems in a Java Environment**", pp. 212-219. In: Fanelli A.M., Pedrycz W., Petrosino A. (eds) Fuzzy Logic and Applications. WILF 2011. Lecture Notes in Computer Science, vol 6857. Springer, Berlin, Heidelberg, https://doi.org/10.1007/978-3-642-23713-3_27

GUAJE

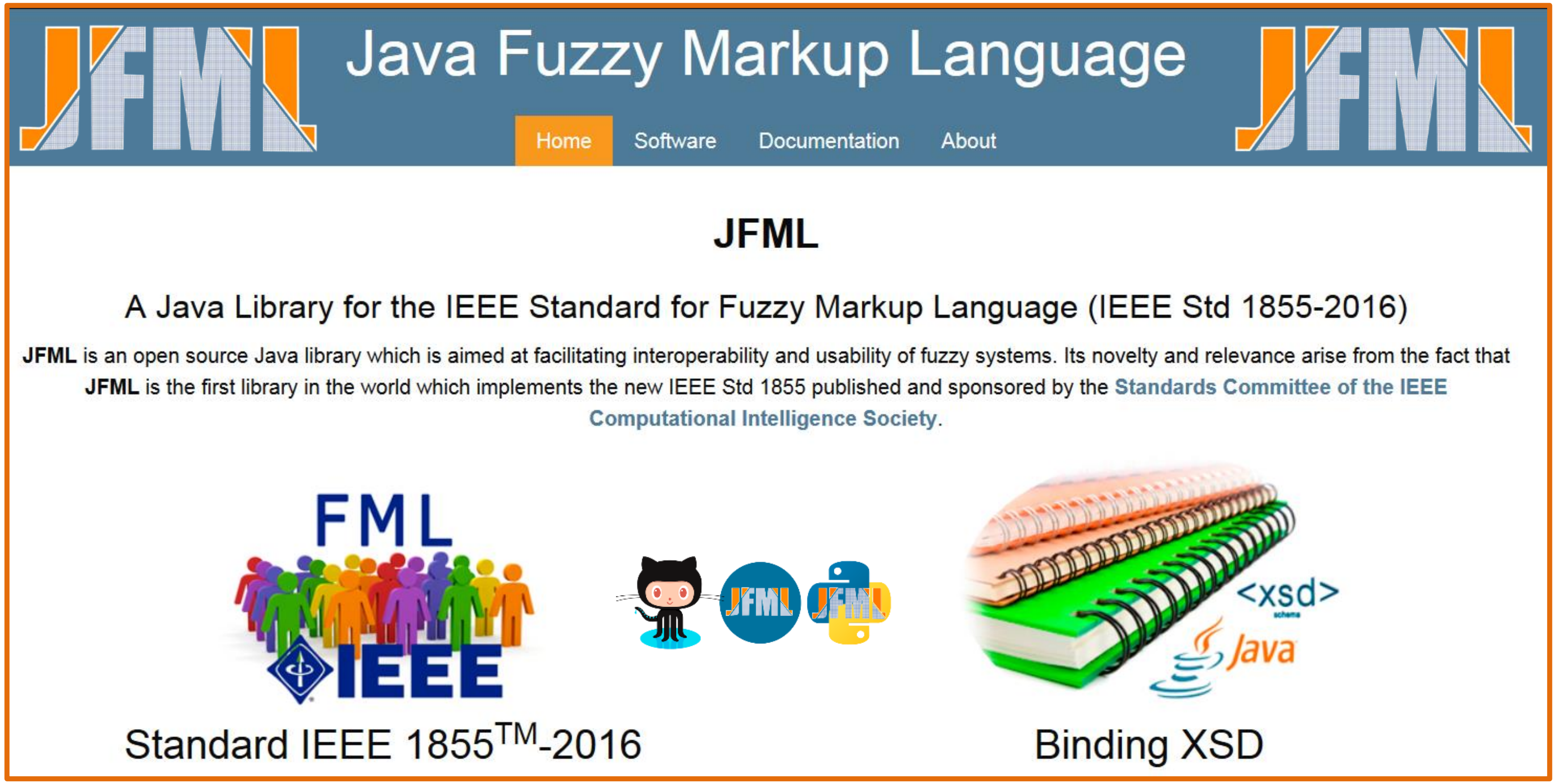
Expert + Data: Fuzzy Rule-based Systems with Global Semantics





Jose M. Alonso, L. Magdalena, "**HILK++: An interpretability-guided fuzzy modeling methodology for learning readable and comprehensible fuzzy rule-based classifiers**", Soft Computing, 15:1959-1980, 2011, https://dx.doi.org/10.1007/s00500-010-0628-5

JFML



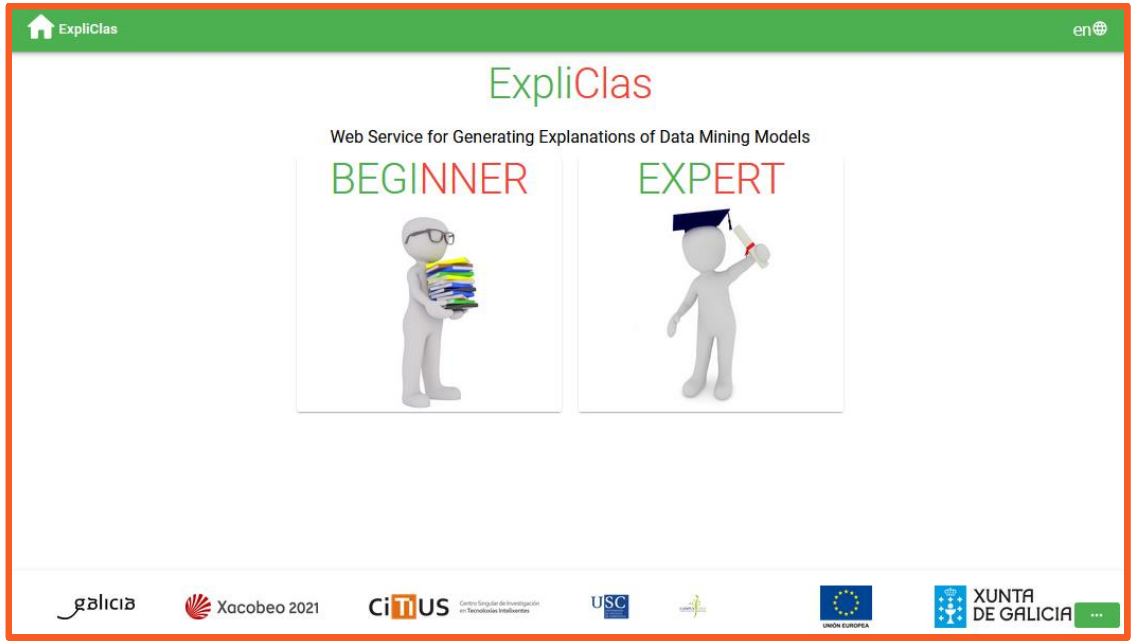
J. M. Soto-Hidalgo, Jose M. Alonso, G. Acampora, J. Alcala-Fdez, "**JFML: A Java Library to Design Fuzzy Logic Systems According to the IEEE Std 1855-2016**," IEEE Access, 6:54952–54964, 2018, https://dx.doi.org/10.1109/ACCESS.2018.2872777

http://www.uco.es/JFML/

ExpliClas

- Weka Classifiers:
  - Black-box Ensemble: **Random Forest**
  - Interpretable Classifiers:
    - Crisp DT: **J48**, **REPTree**, **RandomTree**
    - Fuzzy DT: **FHDT**
    - Fuzzy Rules: **FURIA**
- Textual + Visual Explanations
  - **Textual**: Natural Language (simpleNLG)
  - **Visual**: Trees, Rules, FINGRAMS
- Global + Local Explanations
  - **Global**: Confusion Matrix
  - **Local**: Classification of test instance
- IEEE Std. 1855-2016

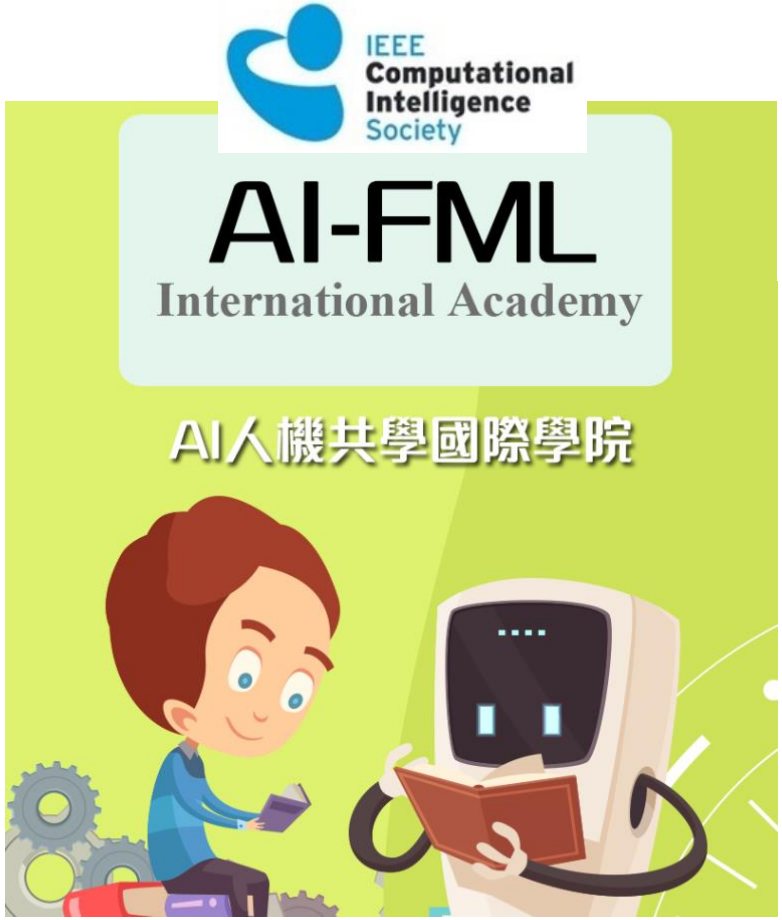
https://demos.citius.usc.es/ExpliClas/
https://demos.citius.usc.es/ExpliClasAPI/
https://www.youtube.com/watch?v=7jfYYudPUZY



Jose M. Alonso, A. Bugarín, "**ExpliClas: Automatic Generation of Explanations in Natural Language for Weka Classifiers**", IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), New Orleans, USA, 1-6, 2019, https://dx.doi.org/10.1109/FUZZ-IEEE.2019.8859018

Use Cases

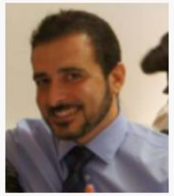
Teaching XAI to Young Students




Hung-Duen Yang
Taiwan


Chang-Shing Lee
Taiwan

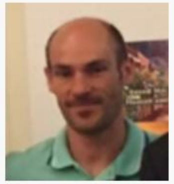

Marek Reformat
Canada


Giovanni Acampora
Italy


Mitsunori Matsushita
Japan


Ryosuke Yamanishi
Japan


Yoko Nishihara
Japan


Jose M. Alonso
Spain


Toru Yamaguchi
Japan


Jose M. Soto Hidalgo
Spain


Kenji Higashi
Japan


Toshitaka Umemoto
Japan


Kiyoshi Hayakawa
Japan


Po-Hsun Cheng
Taiwan


Yusuke Nojima
Japan


Naoyuki Kubota
Japan


Naoki Masuyama
Japan


Ryosuke Saga
Japan


Marie-Jeanne Lesot
France


Amir Pourabdollah
UK

https://sites.google.com/asap.nutn.edu.tw/ai-fml-international-academy/home?authuser=0

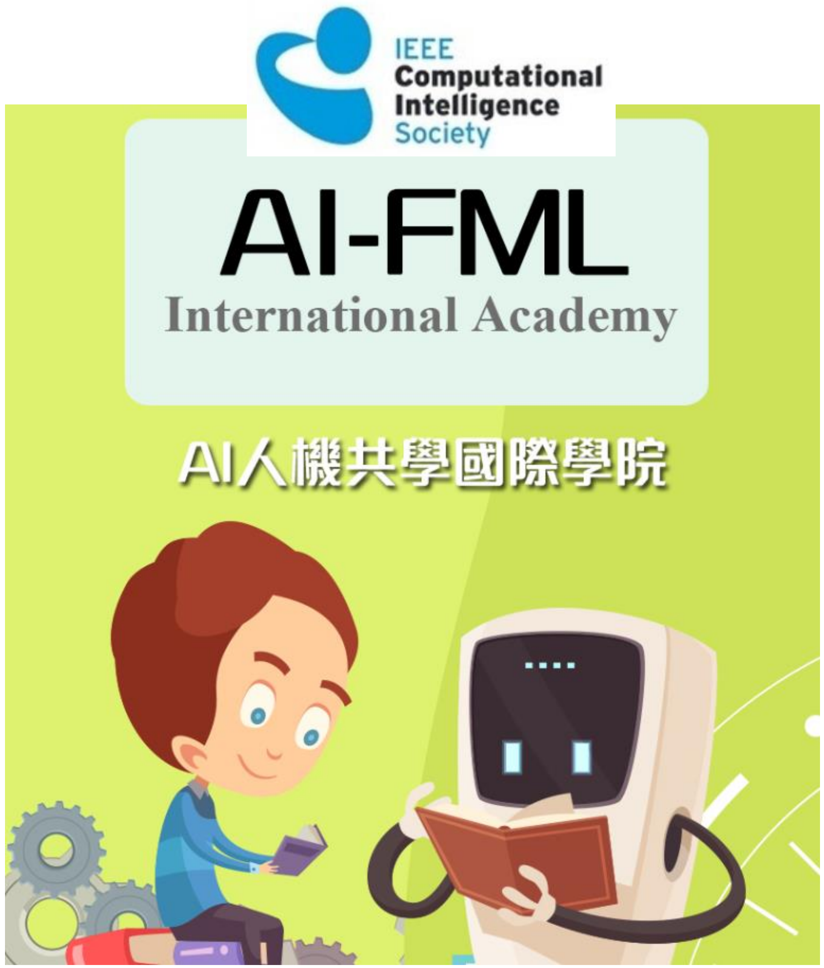
MOST 科技部 Ministry of Science and Technology 數位經濟技術創新研發與應用推動計畫
NUTN
人工智慧產學研聯盟
K.W.S
TOKYO METROPOLITAN UNIVERSITY Research Center
WEBDUINO NUWA 劍湖山世界 長春企管公司

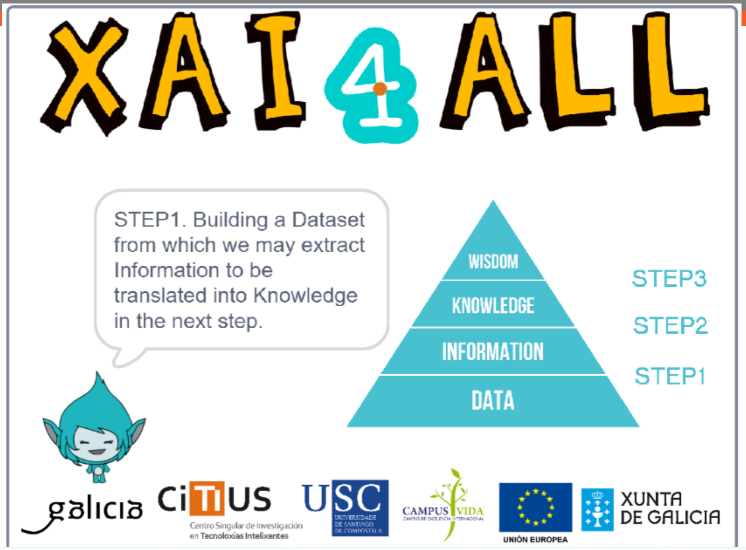
USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
CAMPUS VIDA CAMPUS DE EXCELENCIA INTERNACIONAL
CiTIUS Centro Singular de Investigación en Tecnoloxías Intelixentes

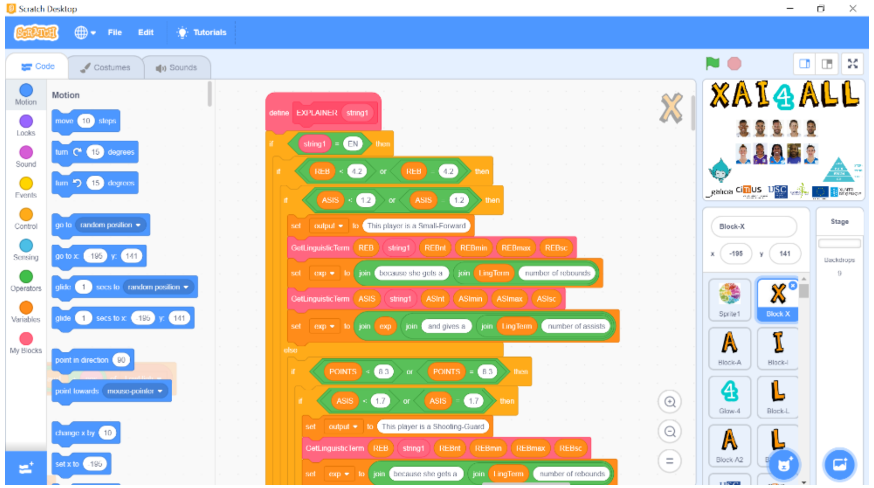
Jose M. Alonso
https://citius.usc.es/v/jose-maria-alonso-moral

SCIS&ISIS2020

Teaching XAI to Young Students







https://sites.google.com/asap.nutn.edu.tw/ai-fml-international-academy/home?authuser=0

https://www.youtube.com/watch?v=7jfYYudPUZY
https://scratch.mit.edu/projects/330579801/

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA · CAMPUS VIDA · CiTIUS Centro Singular de Investigación en Tecnoloxías Intelixentes

Jose M. Alonso
https://citius.usc.es/v/jose-maria-alonso-moral

SCIS&ISIS2020

Teaching XAI to Young Students

C:\GUAJE-vb3.0\tmp\tempexample.jfml.guaje.xml.kb.xml.html

Print Export Close

Decision

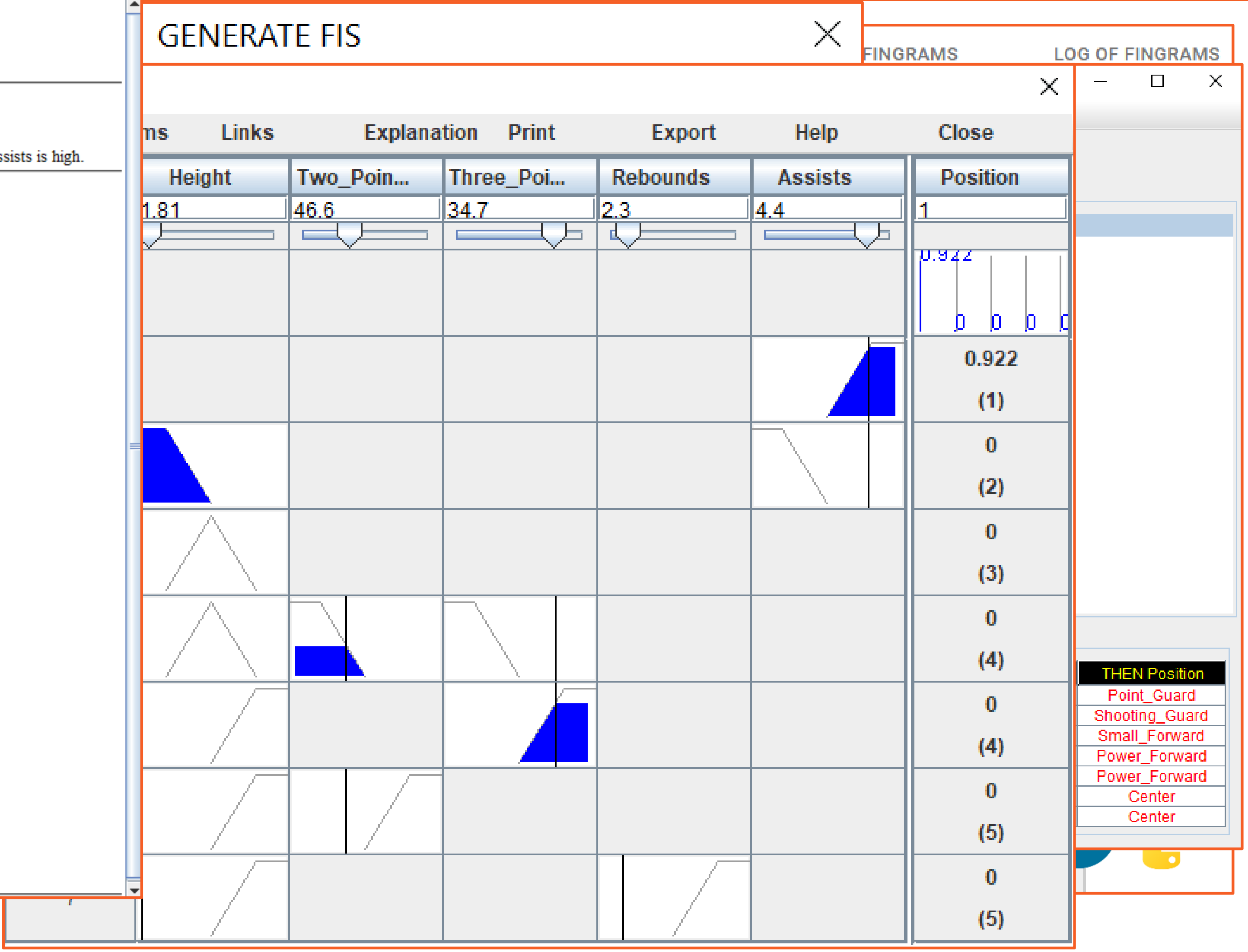
The Position is Point_Guard.

Explanation

We have very high confidence in the classification result. It is very likely that this Position is Point_Guard, because in accordance with rule 1, assists is high.

Additional Details

- **Height:**
  - The height is low because height equals 1.81
  - Height can take values from 1.81 to 2.2
  - Linguistic terms= [Low, Medium, High]
- **Two_Points_Field_Goals_Percentage:**
  - The two_points_field_goals_percentage is medium because two_points_field_goals_percentage equals 46.6
  - Two_Points_Field_Goals_Percentage can take values from 34.4 to 67.7
  - Linguistic terms= [Low, Medium, High]
- **Three_Points_Field_Goals_Percentage:**
  - The three_points_field_goals_percentage is high because three_points_field_goals_percentage equals 34.7
  - Three_Points_Field_Goals_Percentage can take values from 0.0 to 45.5
  - Linguistic terms= [Low, Medium, High]
- **Rebounds:**
  - The rebounds is low because rebounds equals 2.3
  - Rebounds can take values from 1.6 to 6.8
  - Linguistic terms= [Low, Medium, High]
- **Assists:**
  - The assists is high because assists equals 4.4
  - Assists can take values from 0.2 to 5.4
  - Linguistic terms= [Low, Medium, High]



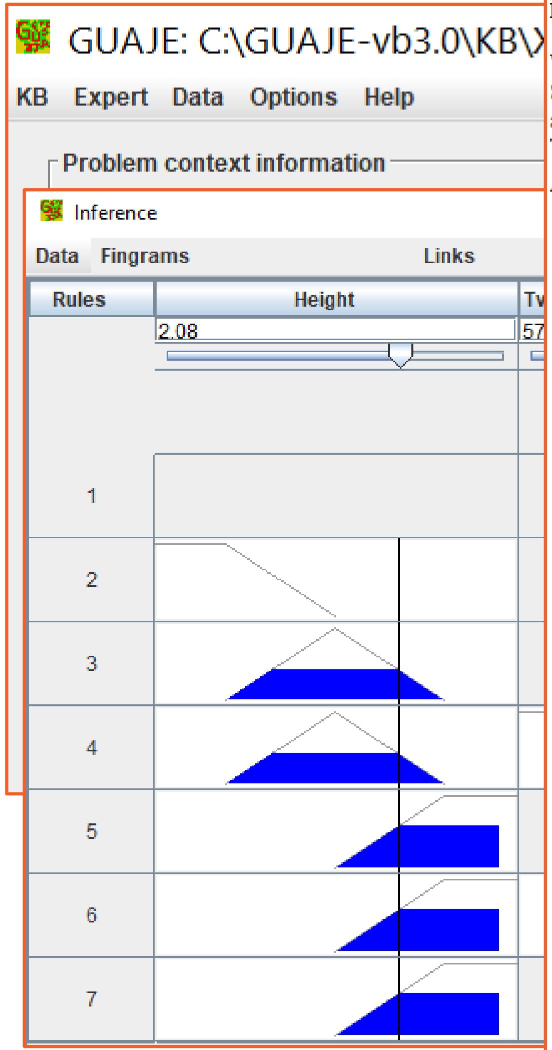
Centro Singular de Investigación en Tecnoloxías Intelixentes

Jose M. Alonso

https://citius.usc.es/v/jose-maria-alonso-moral

Teaching XAI

Scratch + ExpliCl



C:\GUAJE-vb3.0\tmp\tempexample.jfml.guaje.xml.kb.xml.html

Print Export Close

Decision

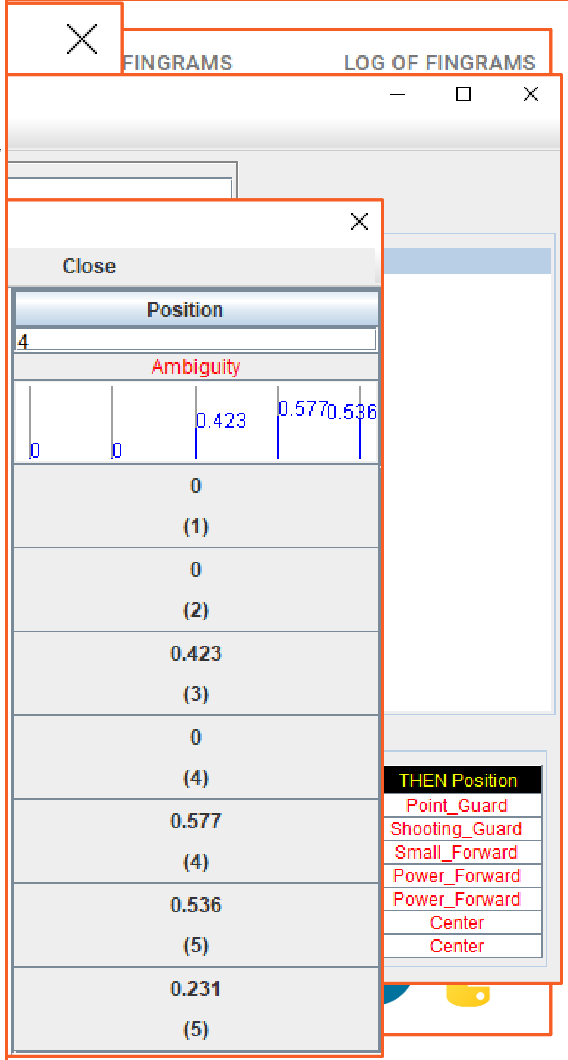
The Position is Power_Forward.

Explanation

We have medium confidence in the classification result. The Position is probably Power_Forward or Center. There is also a smaller chance that it is Small_Forward. On balance, Power_Forward is more likely, because in accordance with rule 5, height and three_points_field_goals_percentage are high. In addition, two_points_field_goals_percentage takes a borderline value between medium and high.

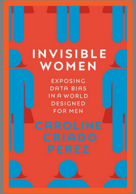
Additional Details

- **Height:**
  - The height is high because height equals 2.08
  - Height can take values from 1.81 to 2.2
  - Linguistic terms= [Low, Medium, High]
- **Two_Points_Field_Goals_Percentage:**
  - The two_points_field_goals_percentage is high because two_points_field_goals_percentage equals 57
  - Two_Points_Field_Goals_Percentage can take values from 34.4 to 67.7
  - Linguistic terms= [Low, Medium, High]
- **Three_Points_Field_Goals_Percentage:**
  - The three_points_field_goals_percentage is high because three_points_field_goals_percentage equals 37.4
  - Three_Points_Field_Goals_Percentage can take values from 0.0 to 45.5
  - Linguistic terms= [Low, Medium, High]
- **Rebounds:**
  - The rebounds is medium because rebounds equals 4.6
  - Rebounds can take values from 1.6 to 6.8
  - Linguistic terms= [Low, Medium, High]
- **Assists:**
  - The assists is low because assists equals 0.9
  - Assists can take values from 0.2 to 5.4
  - Linguistic terms= [Low, Medium, High]



Teaching XAI to Young Students

Data Bias



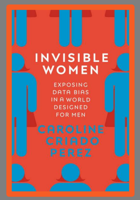
Attribute	MEN	WOMEN	BLACK	WHITE	ALL	Mean
Height	**100**	**87.5**	62.5	**75**	**75**	**80**
Minutes	25	37.5	37.5	50	50	40
Points	25	50	12.5	37.5	37.5	32.5
2P	62.5	37.5	25	50	50	45
3P	37.5	37.5	**75**	50	62.5	52.5
1P	25	37.5	12.5	50	37.5	32.5
Rebounds	62.5	62.5	12.5	62.5	50	50
Assists	50	50	37.5	50	62.5	50
Blocks	12.5	12.5	12.5	**75**	50	32.5
Turnovers	25	37.5	50	37.5	50	40
GA	12.5	37.5	37.5	37.5	**75**	40

	BLACK	WHITE	ALL
MEN	12	38	50
WOMEN	18	32	50
ALL	30	70	100

	MEN	WOMEN	BLACK	WHITE	ALL
RF	70.00	58.00	46.67	**55.71**	**63.00**
J48	58.00	**60.00**	33.33	42.86	55.00
REPT	62.00	52.00	23.33	51.43	46.00
RANDT	54.00	58.00	33.33	47.14	54.00
FHDT3	58.00	46.00	50.00	44.29	39.00
FHDT5	68.00	48.00	40.00	28.57	43.00
FHDT7	**74.00**	40.00	**56.67**	35.71	46.00
FURIA	68.00	48.00	33.33	41.43	53.00

Teaching XAI to Young Students

Data Bias

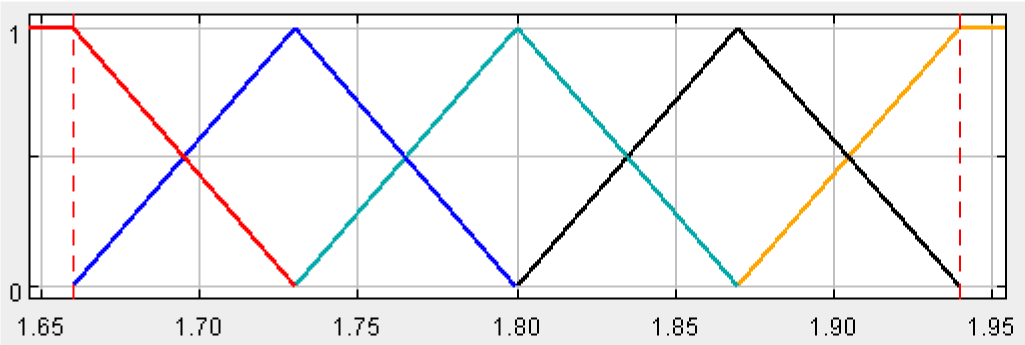


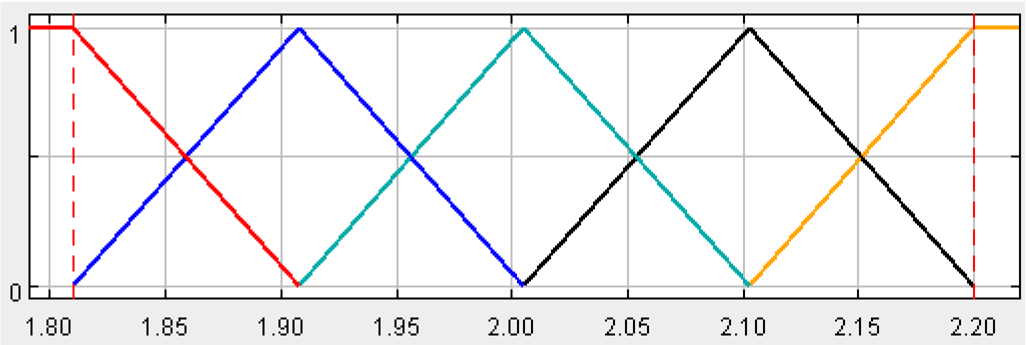
$\text{Height}_{WOMEN} \in [1.66, 1.94]$

- Short = [1/1.66, 0/1.73]
- Medium-height = [0/1.66, 1/1.73, 0/1.8]
- Tall = [0/1.73, 1/1.8, 0/1.87]
- Very tall = [0/1.8, 1/1.87, 0/1.94]
- Extremely tall = [0/1.87, 1/1.94]

$\text{Height}_{MEN} \in [1.81, 2.2]$

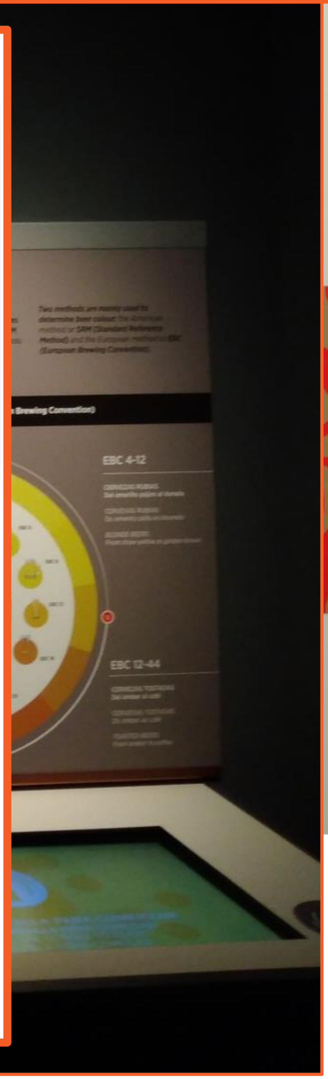
- Short = [1/1.81, 0/1.908]
- Medium-height = [0/1.81, 1/1.908, 0/2.005]
- Tall = [0/1.908, 1/2.005, 0/2.103]
- Very tall = [0/2.005, 1/2.103, 0/2.2]
- Extremely tall = [0/2.103, 1/2.2]

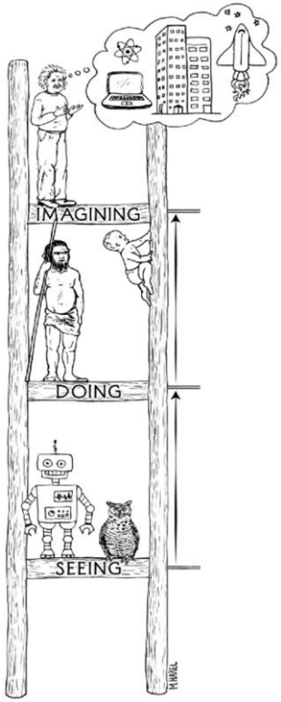


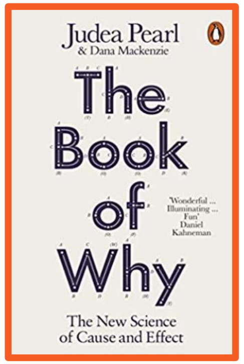


Counterfactuals

A use case on Beer Style Classification







USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
CAMPUS VIDA
CiTIUS Centro Singular de Investigación en Tecnoloxías Intelixentes
Jose M. Alonso
https://citius.usc.es/v/jose-maria-alonso-moral

Counterfactuals

A use case on Beer Style Classification

- **400 instances**
- **8 classes**
  - Blanche, Lager, Pilsner, IPA, Stout, Barleywine, Porter, Belgian Strong Ale
- **3 attributes**
  - **Color** [0, 45]: Pale, Straw, Amber, Brown, Black
  - **Bitterness** [8, 250]: Low, Low-medium, Medium-high, High
  - **Strength** [0.039, 0.136]: Session, Standard, High, Very high
- **Decision Trees**
  - **Crisp DT**: J48, REPTree, RandomTree (WEKA)
  - **FRBC**: FURIA (WEKA), HILK (GUAJE)
- **Linguistic Approximation**
  - Global Semantics via Strong Fuzzy Partitions
  - Linguistic IF-THEN Mamdani rules

C:\GUAJE-vb3.0\KB\BEER\BEER3.txt.aux

Close

items	Active	Color	Bitterness	Strength	Class
1	✔	2.0	18.0	0.049	1.0
2	✔	3.0	13.0	0.045	1.0
3	✔	4.0	17.0	0.052	1.0
4	✔	3.0	19.0	0.051	1.0
5	✔	3.0	14.0	0.047	1.0
6	✔	3.0	14.0	0.05	1.0
7	✔	3.0	11.0	0.046	1.0
8	✔	4.0	14.0	0.051	1.0
9	✔	4.0	15.0	0.051	1.0
10	✔	3.0	19.0	0.053	1.0
11	✔	3.0	10.0	0.048	1.0
12	✔	3.0	8.0	0.052	1.0
13	✔	2.0	17.0	0.05	1.0
14	✔	3.0	14.0	0.048	1.0
15	✔	3.0	14.0	0.044	1.0
16	✔	3.0	18.0	0.052	1.0
17	✔	3.0	10.0	0.046	1.0
18	✔	2.0	16.0	0.051	1.0
19	✔	4.0	12.0	0.048	1.0
20	✔	3.0	15.0	0.047	1.0
21	✔	2.0	11.0	0.049	1.0
22	✔	4.0	11.0	0.053	1.0
23	✔	2.0	17.0	0.046	1.0
24	✔	2.0	16.0	0.047	1.0
25	✔	3.0	15.0	0.048	1.0
26	✔	4.0	14.0	0.046	1.0
27	✔	2.0	17.0	0.049	1.0
28	✔	3.0	15.0	0.051	1.0
29	✔	4.0	11.0	0.045	1.0
30	✔	3.0	12.0	0.054	1.0
31	✔	3.0	23.0	0.052	1.0
32	✔	2.0	19.0	0.053	1.0
33	✔	0.0	16.0	0.043	1.0
34	✔	3.0	17.0	0.06	1.0
35	✔	0.0	16.0	0.046	1.0
36	✔	3.0	16.0	0.054	1.0
37	✔	3.0	22.0	0.05	1.0
38	✔	3.0	15.0	0.047	1.0
39	✔	4.0	21.0	0.052	1.0
40	✔	3.0	10.0	0.054	1.0
41	✔	3.0	12.0	0.048	1.0

Jose M. Alonso
https://citius.usc.es/v/jose-maria-alonso-moral

Counterfactuals

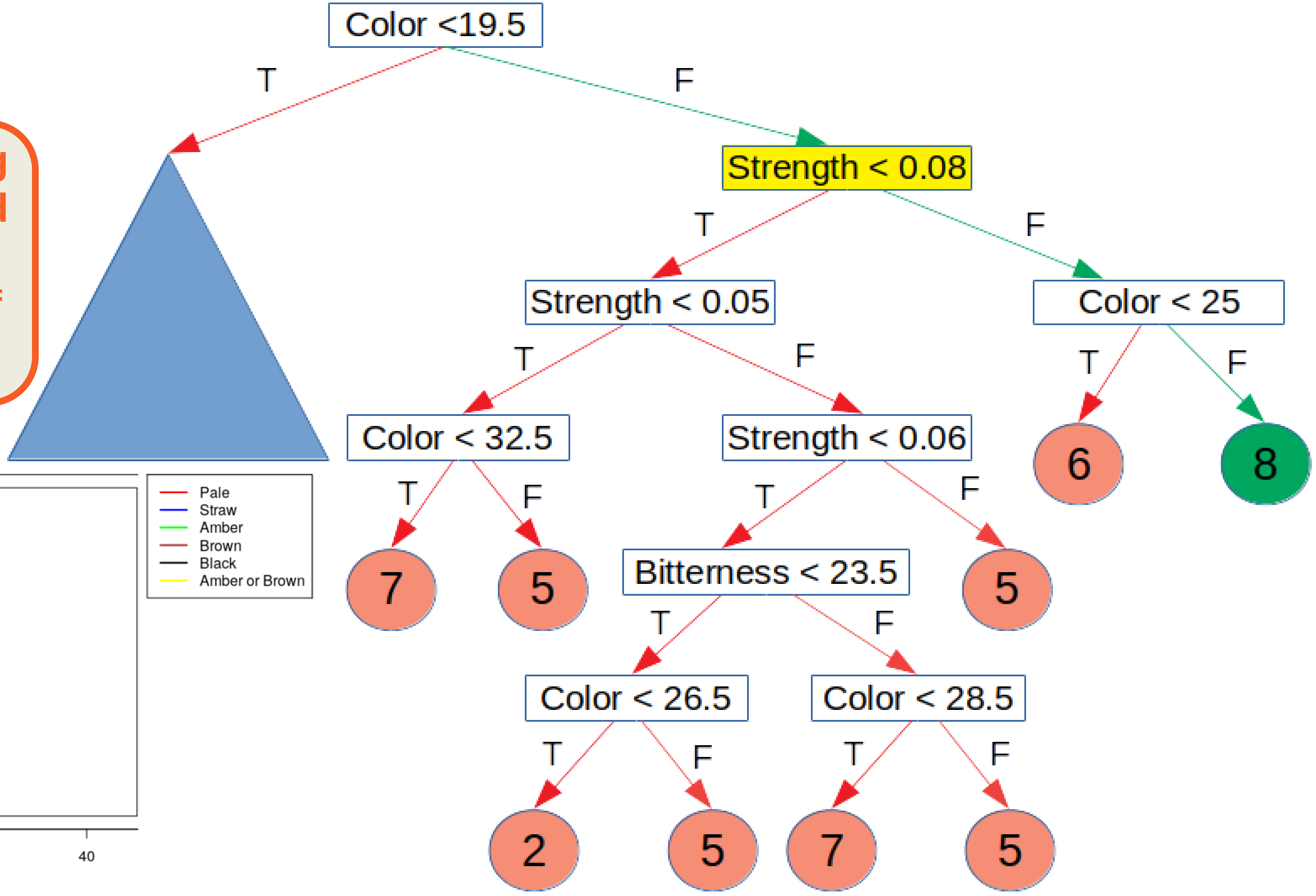
A use case on Beer Style Classification

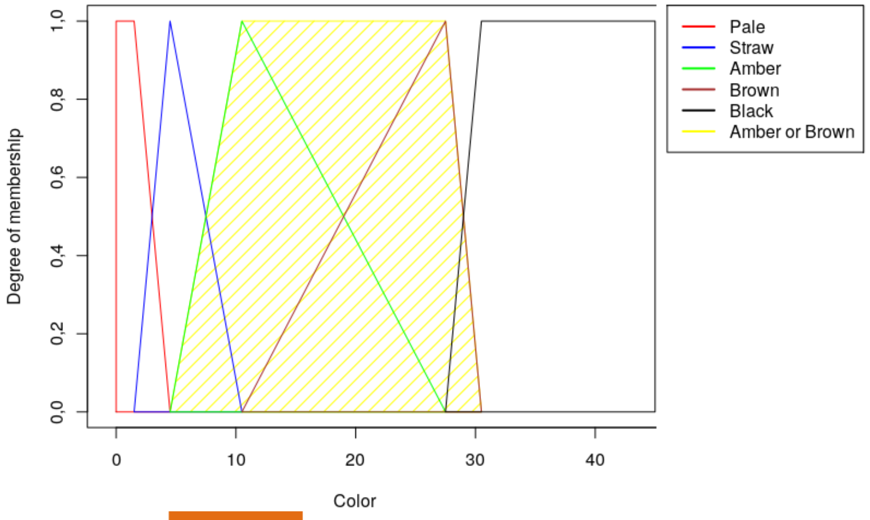
Algorithm	Accuracy (%)	Number of Rules	Total Rule Length
Random Forest	96.25	-	-
J48	95.00	9.8	23.4
FURIA	95.50	14.2	44.0
HILK	93.63	14.4	32.2

Counterfactuals

A use case on Beer Style Classification

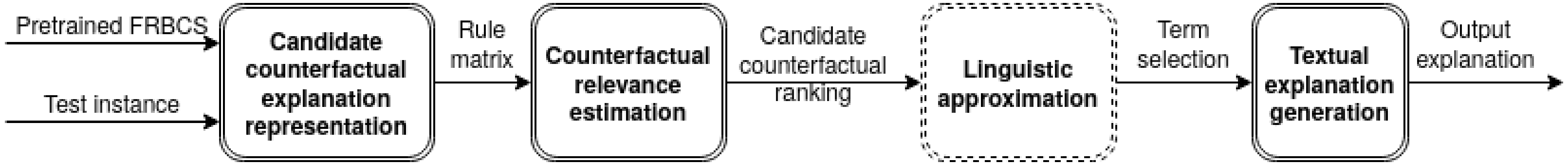
Output class is Belgian Strong Ale because color is black and strength is high. However, output class would be Stout if strength were standard.





Counterfactuals

A use case on Beer Style Classification




```
(Bitterness in [-inf, -inf, 19, 30]) and (Color in [-inf, -inf, 4, 7])
 => Beer_Style=1.0 (CF = 0.97)

(Color in [-inf, -inf, 4, 7]) and (Bitterness in [-inf, -inf, 23, 30])
 => Beer_Style=1.0 (CF = 0.97)

(Bitterness in [-inf, -inf, 27, 52]) and (Color in [6, 7, inf, inf])
 and (Color in [-inf, -inf, 14, 16]) => Beer_Style=2.0 (CF = 0.95)

(Color in [-inf, -inf, 6, 8]) and (Bitterness in [29, 30, inf, inf])
 and (Bitterness in [-inf, -inf, 52, 63]) => Beer_Style=3.0 (CF = 0.97)

(Bitterness in [52, 57, inf, inf])
 and (Strength in [-inf, -inf, 0.091, 0.092]) => Beer_Style=4.0 (CF = 0.89)

(Bitterness in [48, 50, inf, inf]) and (Color in [-inf, -inf, 8, 10])
 => Beer_Style=4.0 (CF = 0.84)

(Color in [30, 31, inf, inf]) => Beer_Style=5.0 (CF = 0.96)

(Color in [29, 30, inf, inf]) and (Bitterness in [-inf, -inf, 32, 39])
 and (Strength in [0.046, 0.053, inf, inf]) => Beer_Style=5.0 (CF = 0.96)

(Strength in [0.087, 0.093, inf, inf])
 and (Bitterness in [47, 50, inf, inf]) and (Color in [11, 12, inf, inf])
 => Beer_Style=6.0 (CF = 0.96)

(Bitterness in [93, 95, inf, inf])
 and (Strength in [0.089, 0.092, inf, inf]) > Beer_Style=6.0 (CF = 0.88)

(Bitterness in [41, 77, inf, inf])
 and (Strength in [0.096, 0.097, inf, inf]) => Beer_Style=6.0 (CF = 0.95)

(Bitterness in [62, 79, inf, inf]) and (Color in [13, 14, inf, inf])
 and (Bitterness in [-inf, -inf, 85, 98]) => Beer_Style=6.0 (CF = 0.9)

(Color in [12, 20, inf, inf]) and (Strength in [-inf, -inf, 0.051, 0.053])
 and (Color in [-inf, -inf, 30, 34]) => Beer_Style=7.0 (CF = 0.96)

(Color in [17, 20, inf, inf]) and (Strength in [-inf, -inf, 0.054, 0.082])
 and (Color in [-inf, -inf, 29, 30]) => Beer_Style=7.0 (CF = 0.95)

(Strength in [0.07, 0.087, inf, inf])
 and (Bitterness in [-inf, -inf, 47, 50]) => Beer_Style=8.0 (CF = 0.97)
```

Feature	*Linguistic term*	*Range of values*
Color	Pale	0 ... 3
	Straw	3 ... 7.5
	Amber	7.5 ... 19
	Brown	19 ... 29
	Black	29 ... 45
Bitterness	Low	7 ... 21
	Low-medium	21 ... 32.5
	Medium-high	32.5 ... 47.5
	High	47.5 ... 250
Strength	Session	0.035 ... 0.0525
	Standard	0.0525 ... 0.0675
	High	0.0675 ... 0.09
	Very high	0.09 ... 0.136

C:\GUAJE-vb3.0\tmp\tempBEER3.txt..FURIA.jfml.kb.xml.html

Print Export Close

Decision

The Beer_Style is Porter.

Explanation

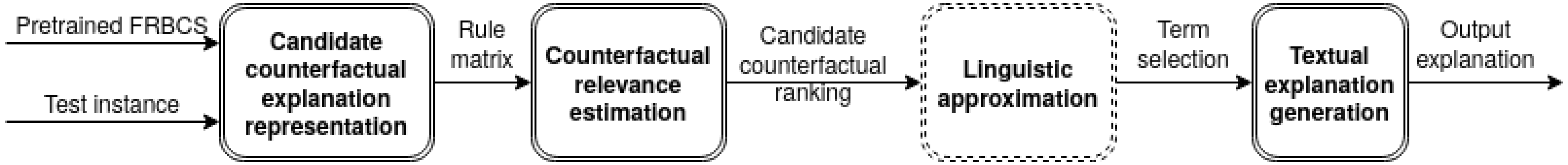
We have medium confidence in the classification result. The Beer_Style is probably Porter. There is also a small chance that it is Stout. On balance, Porter is more likely, because in accordance with rule 13, color is brown and strength is session.

Additional Details

- **Color:**
  - The color is brown because color equals 29
  - Color can take values from 0.0 to 45.0
  - Linguistic terms= [Pale, Straw, Amber, Brown, Black]
- **Bitterness:**
  - The bitterness is low-medium because bitterness equals 26
  - Bitterness can take values from 8.0 to 250.0
  - Linguistic terms= [Low, Low-medium, Medium-high, High]
- **Strength:**
  - The strength is session because strength equals 0.046
  - Strength can take values from 0.039 to 1.136
  - Linguistic terms= [Session, Standard, High, Very-high]

Counterfactuals

A use case on Beer Style Classification



$$x =< (\text{Color}, 12.0); (\text{Bitterness}, 104.0); (\text{Strength}, 0.099) : (\text{Beer}, \text{Barleywine}) >$$

Rule number	*Activation score*	*Predicted class*
3	0.156	Pilsner
4	0.156	IPA
5	0.409	IPA
8	**0.589**	Barleywine
12	0.411	Belgian Strong Ale

Output class is Barleywine because bitterness is high and strength is very high. However, output class would be IPA if strength were high.

$$e_f(x,s) = \{(\text{Bitterness}, \text{High}); (\text{Strength}, \text{Very high}) : (\text{Beer}, \text{Barleywine})\}$$

$$e_{cf}(x,s,IPA) = \{(\text{Bitterness}, \text{High}); (\text{Strength}, \text{High}) : (\text{Beer}, \text{IPA})\}$$

$$\boxed{E(x,s) = e_f \cup E_{cf} = \{(\text{Bitterness}, \text{High}); (\text{Strength}, \text{Very high}) : (\text{Beer}, \text{Barleywine}); \{(\text{Bitterness}, \text{High}); (\text{Strength}, \text{High}) : (\text{Beer}, \text{IPA})\}\}}$$

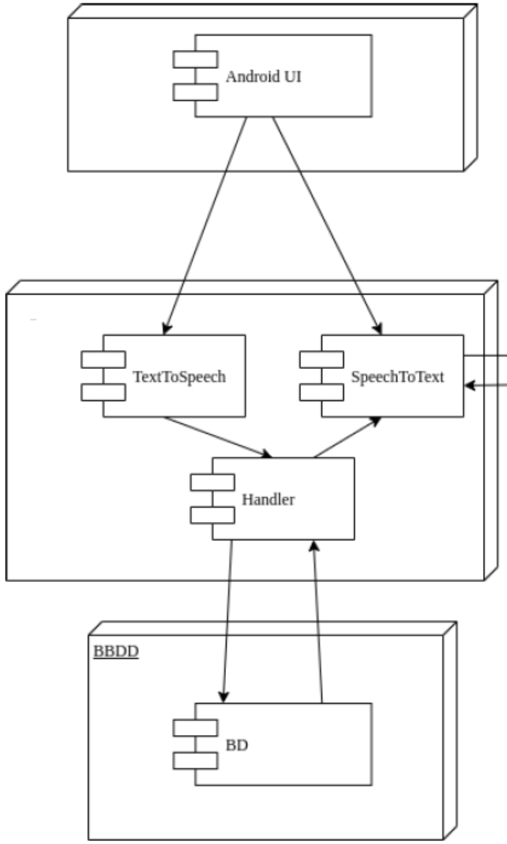
XAI for e-Health

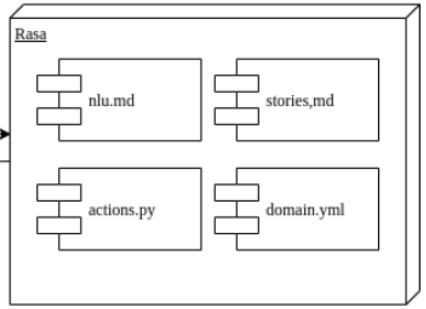
ADHERE-U Project (RTI2018-099646-B-I00)

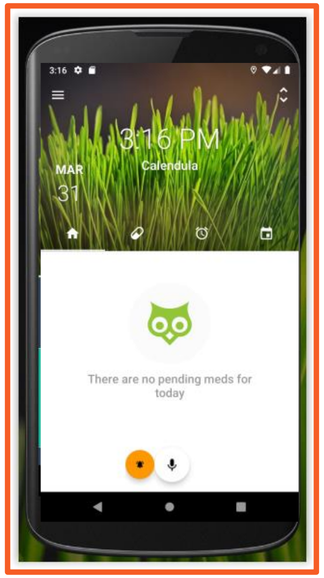
AGENCIA ESTATAL DE INVESTIGACIÓN

GOBIERNO DE ESPAÑA · MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES

- **Models, techniques and methodologies based on AI for improving Medication Adherence**
  - **Data Exploitation Services**: Symptom Patterns Discovery, Emotion Recognition
  - **Persuasive Interaction**: Conversational Interface with Interactive Explanations and Adaptive Communication Strategy
    - **CalendulaBot:** https://gitlab.citius.usc.es/jose.guerra.vilar/calendulabot
  - **Evaluation**: Clinical Trials with Polymedicated Patients; Patients with Lung Cancer; and Patients with Atrial Fibrillation



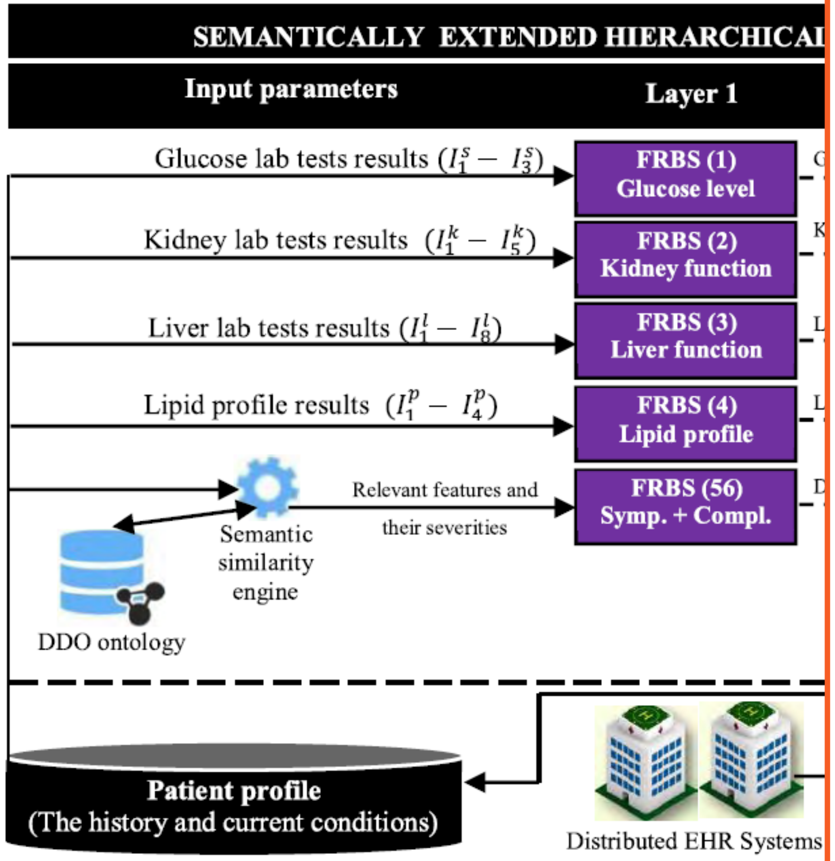


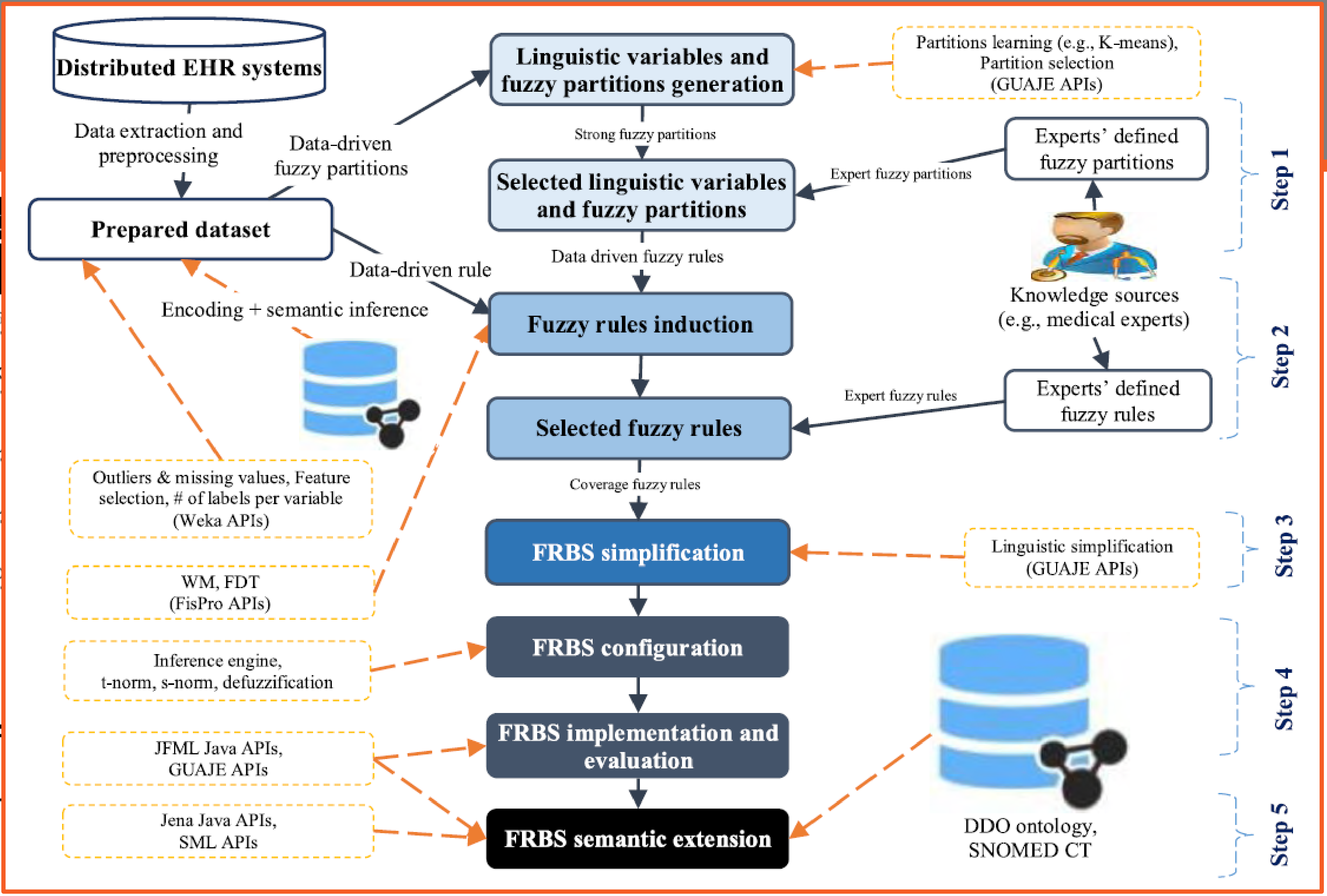


Jose M. Alonso
https://citius.usc.es/v/jose-maria-alonso-moral

XAI for e-Health

Diabetes Diagnosis



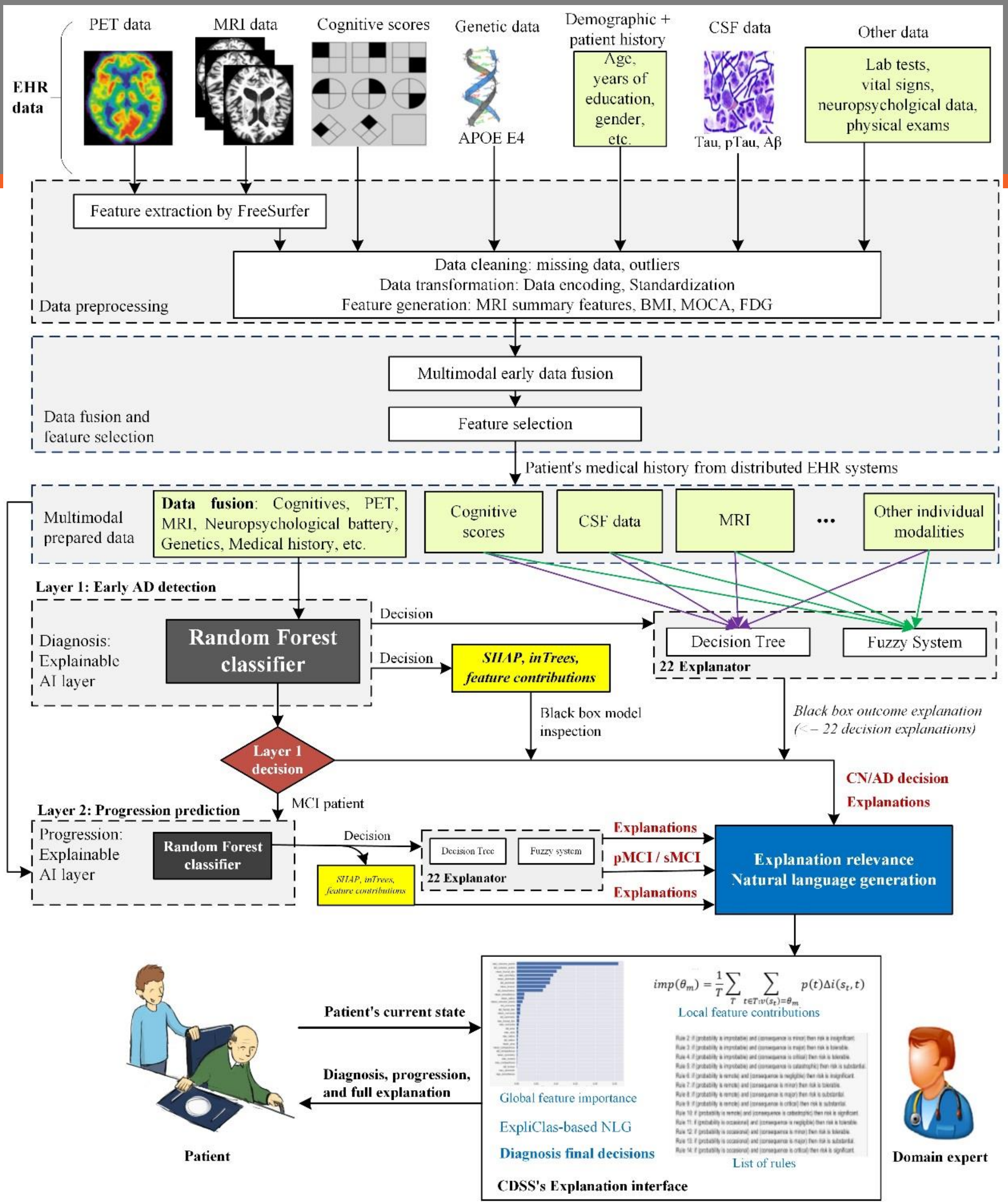


Shaker El-Sappagh, Jose M. Alonso, F. Ali, A. Ali, Jun-Hyeog Jang, Kyung-Sup Kwak, "**An ontology-based interpretable fuzzy decision support system for diabetes diagnosis**", IEEE Access, 6:37371-37394, 2018, https://dx.doi.org/10.1109/ACCESS.2018.2852004

XAI for e-Health

Alzheimer's Disease (AD)

- **A Multilayer Multimodal Detection and Prediction Model based on Explainable Artificial Intelligence for Alzheimer's Disease**
  (under review)



Centro Singular de Investigación en Tecnoloxías Intelixentes

Jose M. Alonso

https://citius.usc.es/v/jose-maria-alonso-moral

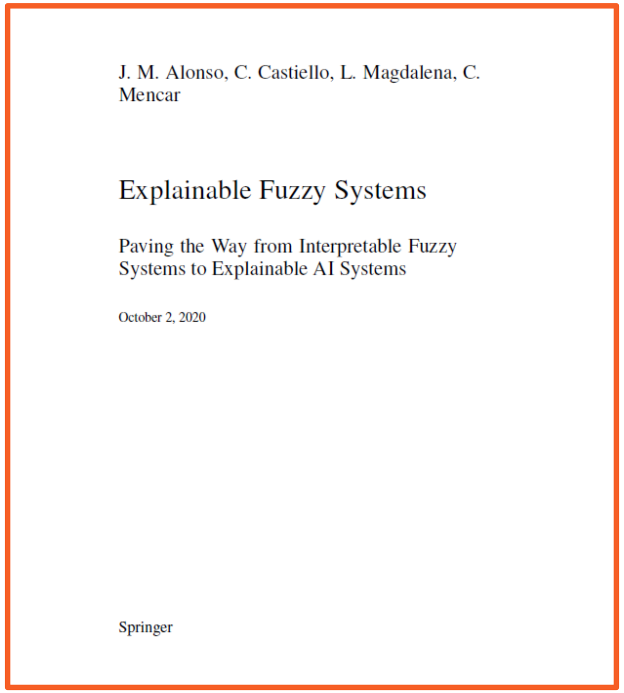
SCIS&ISIS2020

- **A Multilayer Multimodal Detection and Prediction Model based on Explainable Artificial Intelligence for Alzheimer's Disease** (under review)
  - **1048 subjects** from ADNI dataset (http://adni.loni.usc.edu/):
    - Cognitively normal (CN): 294 subjects
    - sMCI: 254 subjects, pMCI: 232 subjects, and AD: 268 subjects
  - **11 modalities**: PET, MRI, Cognitive Scores, Demographic, etc.
  - Two-layer Accurate and Interpretable AD diagnosis and progression detection model
    - First layer: **diagnosis AD patients**
      (accuracy: 93.95% and F1-score: 93.94%)
    - Second layer: **MCI-to-AD progression within three years from baseline diagnosis**
      (accuracy: 87.08% and F1-Score: 87.09%)
    - **Global and instance-based explanations** using SHAP
    - 22 **NL explainers based on decision trees and fuzzy rule-based systems**

ADNI
ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE

References

- Jose M. Alonso, C. Castiello, L. Magdalena, C. Mencar, “**Explainable Fuzzy Systems: Paving the way from Interpretable Fuzzy Systems to Explainable AI Systems**”, Studies in Computational Intelligence, Springer Nature Switzerland AG, In press, 2021
- Jose M. Alonso, “**Teaching Explainable Artificial Intelligence to High School Students**”, International Journal of Computational Intelligence Systems, 13(1):974-987, 2020, https://dx.doi.org/10.2991/ijcis.d.200715.003
- C. Mencar, Jose M. Alonso, "**Paving the way to explainable Artificial Intelligence with fuzzy modeling**", WILF2018 - 12th International Workshop on Fuzzy Logic and Applications, Springer, pp. 215-227, LNCS11291, 2019, https://dx.doi.org/10.1007/978-3-030-12544-8_17
- Jose M. Alonso, A. Bugarín, “**ExpliClas: Automatic Generation of Explanations in Natural Language for Weka Classifiers**”, IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), New Orleans, USA, 1-6, 2019, https://dx.doi.org/10.1109/FUZZ-IEEE.2019.8859018
- Jose M. Alonso, Pietro Ducange, Riccardo Pecori, Raul Vilas, “**Building Explanations for Fuzzy Decision Trees with the ExpliClas Software**”, IEEE World Congress on Computational Intelligence (IEEE-WCCI), Glasgow, Scotland, 2020, https://dx.doi.org/10.1109/FUZZ48607.2020.9177725
- Jose M. Alonso, Javier Toja-Alamancos, Alberto Bugarin, “**Experimental Study on Generating Multi-modal Explanations of Black-box Classifiers in terms of Gray-box Classifiers**”, IEEE World Congress on Computational Intelligence (IEEE-WCCI), Glasgow, Scotland, 2020, https://dx.doi.org/10.1109/FUZZ48607.2020.9177770
- I. Stepin, Jose M. Alonso, A. Catala, M. Pereira-Fariña, “**Generation and Evaluation of Factual and Counterfactual Explanations for Decision Trees and Fuzzy Rule-based Classifiers**”, IEEE World Congress on Computational Intelligence (IEEE-WCCI), Glasgow, Scotland, 2020, https://dx.doi.org/10.1109/FUZZ48607.2020.9177629
- Shaker El-Sappagh, Jose M. Alonso, F. Ali, A. Ali, Jun-Hyeog Jang, Kyung-Sup Kwak, “**An ontology-based interpretable fuzzy decision support system for diabetes diagnosis**”, IEEE Access, 6:37371-37394, 2018, https://dx.doi.org/10.1109/ACCESS.2018.2852004



J. M. Alonso, C. Castiello, L. Magdalena, C. Mencar

Explainable Fuzzy Systems

Paving the Way from Interpretable Fuzzy Systems to Explainable AI Systems

October 2, 2020

Springer

https://gitlab.citius.usc.es/jose.alonso/bookexfs/

https://gitlab.citius.usc.es/jose.alonso/xai

Task Force on Explainable Fuzzy Systems

IEEE Computational Intelligence Society

Fuzzy Systems Technical Committee

The Task Force on Explainable Fuzzy Systems (TF-EXFS) is a Task Force of the Fuzzy Systems Technical Committee (FSTC) which belongs to the Computational Intelligence Society (CIS) of the Institute of Electrical and Electronic Engineers, Inc. (IEEE).

- ✓ The **mission** of this Task Force is to lead the development of a new generation of Explainable Fuzzy Systems, with a holistic view of **fundamentals and current research trends** in the XAI field, paying special attention to fuzzy-grounded knowledge representation and reasoning but also regarding how to enhance human-machine interaction through multi-modal (e.g., graphical or textual modalities) **effective explanations**.
- ✓ Given the multidisciplinary nature of the XAI research field, the scope of this task force goes beyond the usual topics treated by the fuzzy community. The activities to be developed will be of interest for researchers, from both academy and industry, working in the fields of Artificial and Computational Intelligence (with **special attention to Fuzzy Logic** but addressing **also XAI challenges on Neural Networks, Evolutionary Computation, Bayesian Networks, Bio-inspired algorithms**, etc.).

https://sites.google.com/view/tf-explainable-fuzzy-systems/

IEEE Computational Intelligence Society

Special Issue

IEEE

TF-EXFS Explainable and Trustworthy Artificial Intelligence

Journal: IEEE Computational Intelligence Magazine

Editors: Jose M. Alonso, Corrado Mencar, Hisao Ishibuchi

- Paper submission: **February 15th, 2021**
- Acceptance/rejection notification: **April 15th, 2021**
- Revision due: **May 15th, 2021**
- Final notification: **July 1st, 2021**
- Camera-ready due: **July 15th, 2021**
- Publication Date: **November 2021**

IEEE Computational Intelligence MAGAZINE

Impact Score

9.083 Impact Factor	0.00277 Eigenfactor	2.468 Article Influence Score

https://sites.google.com/view/special-issue-on-xai-ieee-cim

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

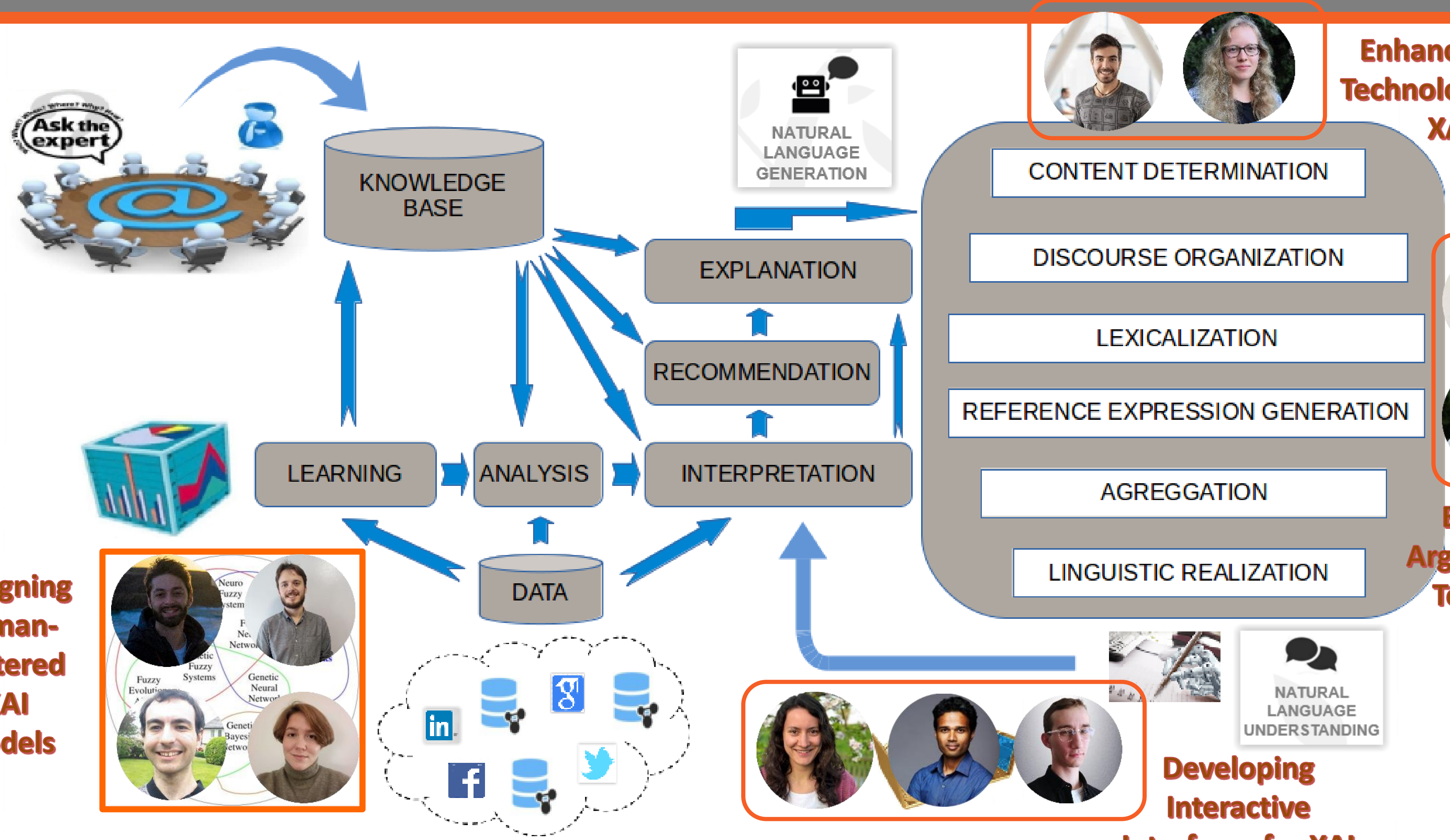
CAMPUS VIDA CAMPUS DE EXCELENCIA INTERNACIONAL

CiTIUS Centro Singular de Investigación en Tecnoloxías Intelixentes

Jose M. Alonso
https://citius.usc.es/v/jose-maria-alonso-moral

SCIS&ISIS2020

NL4XAI: Interactive Natural Language Technology for XAI



Enhancing NL Technologies for XAI

Exploiting Argumentation Technology for XAI

Designing human-centered XAI models

Developing Interactive Interfaces for XAI

NL4XAI

Interactive *Natural Language* Technology for eXplainable Artificial Intelligence

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 860621.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

CAMPUS VIDA CAMPUS DE EXCELENCIA INTERNACIONAL

CiTIUS Centro Singular de Investigación en Tecnoloxías Intelixentes

Jose M. Alonso

https://citius.usc.es/v/jose-maria-alonso-moral

SCIS&ISIS2020

Jose M. Alonso

"Ramón y Cajal" Researcher (RYC-2016-19802) CiTIUS-USC

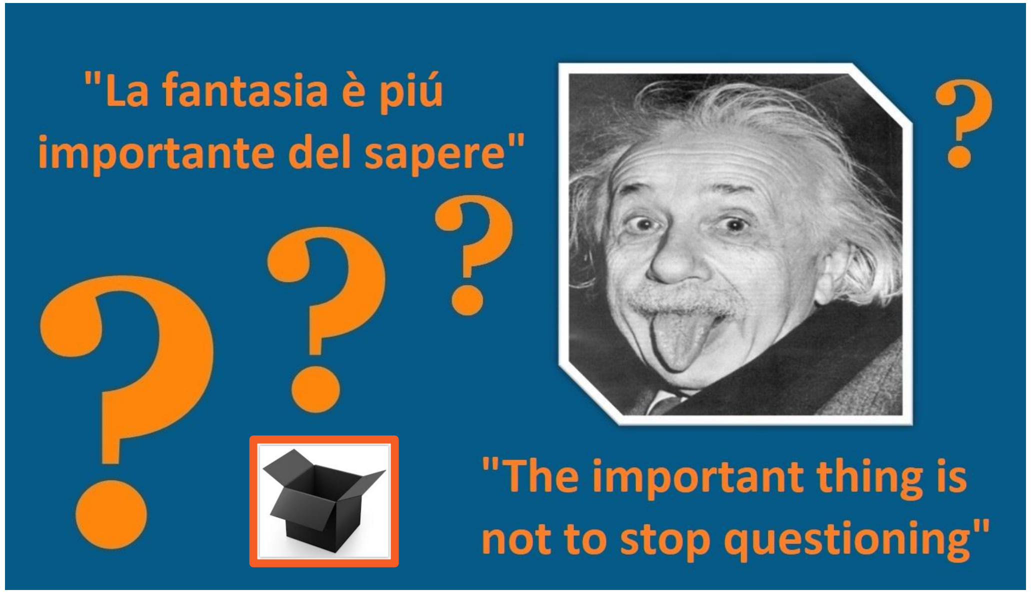
Research Group on Intelligent Systems

josemaria.alonso.moral@usc.es

https://citius.usc.es/v/jose-maria-alonso-moral

https://gitlab.citius.usc.es/jose.alonso/xai

https://demos.citius.usc.es/ExpliClas/



Collaborators

Alberto Bugarín
CiTIUS-USC, Spain

Alejandro Catala
CiTIUS-USC, Spain

Luis Magdalena
Technical University of Madrid, Spain

Corrado Mencar
University of Bari Aldo Moro, Italy

Martín Pereira
USC, Spain

Ciro Castiello
University of Bari Aldo Moro, Italy

Shaker El-Sappagh
CiTIUS-USC, Spain

Ilia Stepin
CiTIUS-USC, Spain

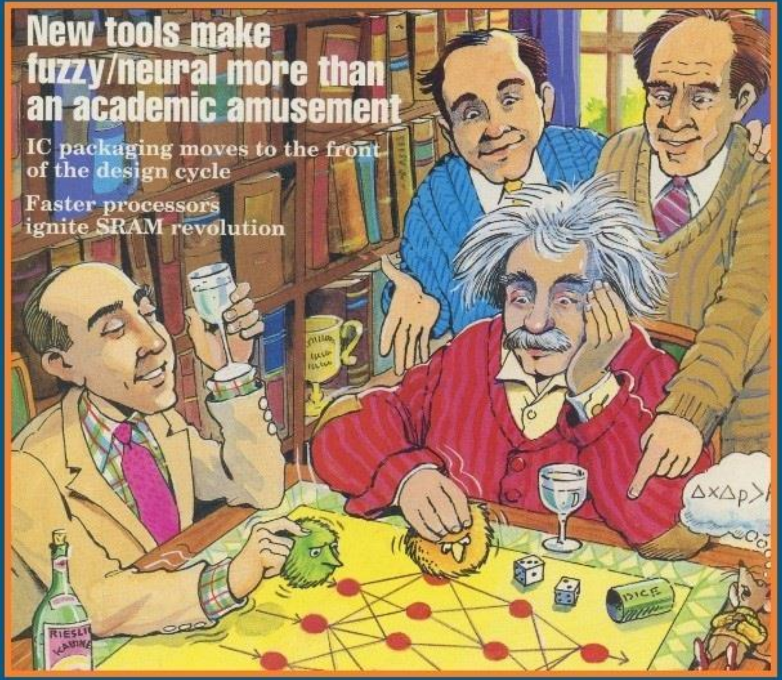
J. M. Soto Hidalgo
University of Granada, Spain

Jesús Alcala-Fdez
University of Granada, Spain

"If there is effort, there is always accomplishment"

"Never be proud of having won an opponent. Who you won today can beat you tomorrow"

"The only victory that endures is the conquest over the own ignorance "



(*) Picture taken from the Fuzzy Archive Rudolf Seising

From left to right: Lotfi A. Zadeh, Niels Bohr, Albert Einstein, and Werner Heisenberg

galicia Xacobeo 2021

UNIÓN EUROPEA

XUNTA DE GALICIA

Fondo Europeo de Desenvolvemento Rexional *"Unha maneira de facer Europa"*

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

CAMPUS VIDA CAMPUS DE EXCELENCIA INTERNACIONAL

CiTIUS Centro Singular de Investigación en Tecnoloxías Intelixentes

Jose M. Alonso
https://citius.usc.es/v/jose-maria-alonso-moral

Interactive Natural Language Technology for Human-Centric Explainable Artificial Intelligence

Jose M. Alonso

Centro Singular de Investigación en Tecnoloxías Intelixentes

Universidade de Santiago de Compostela