🤖 Practical Data Science and Machine Learning 🤖

Juan Cruz-Benito

🦚 @_juancb

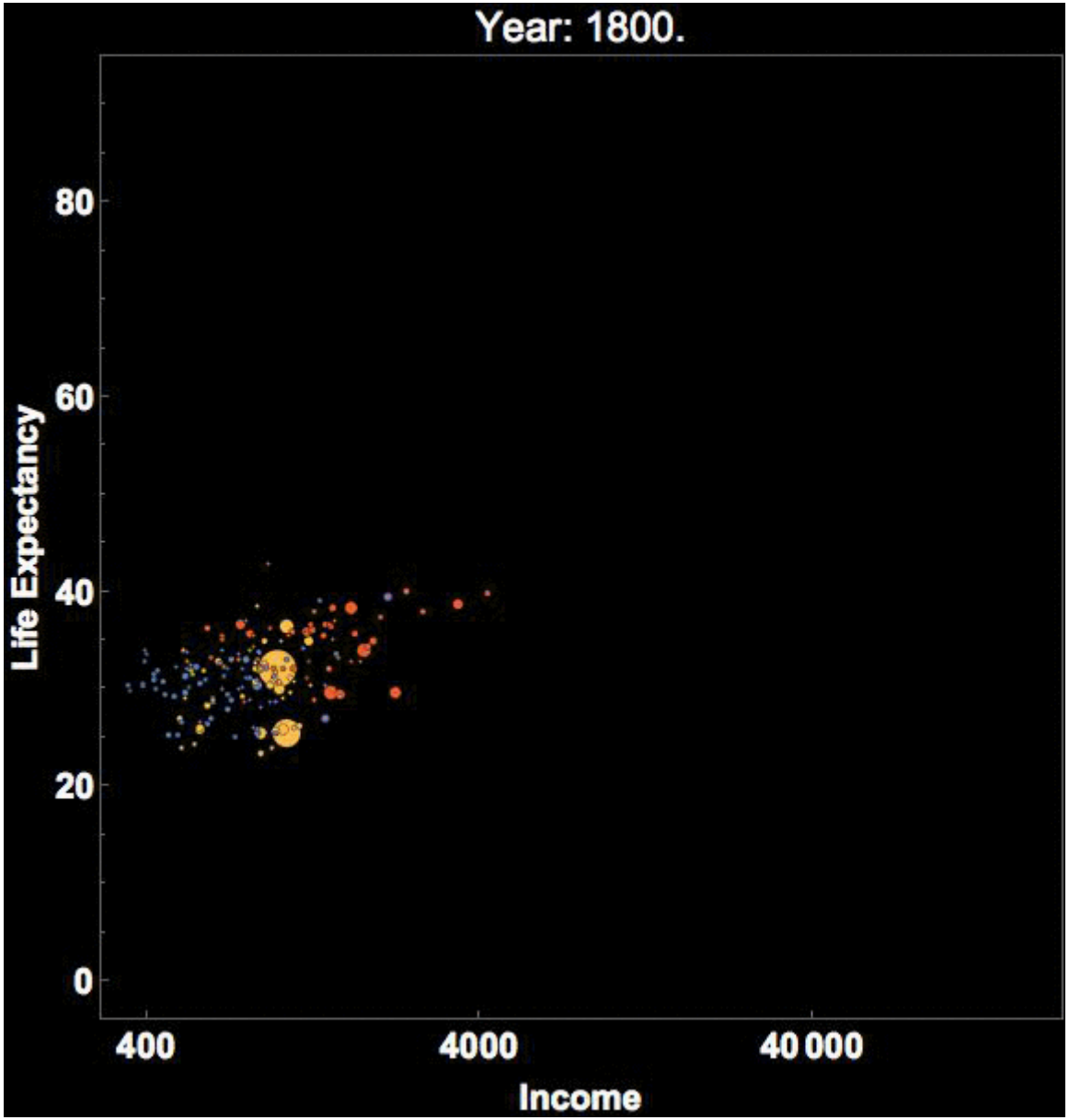
Master on Intelligent Systems. University of Salamanca. 27 Feb 2020

Who am I?

👋 Juan Cruz-Benito

- 👨‍🎓 PhD in Computer Engineering by the University of Salamanca
- 🏆 National award for Spanish young researchers 2019
- 👨‍💻 Senior Software Engineer @ IBM Research Quantum & AI
- 🤓 Data Nerd

What I'm doing here?


Year: 1800.
Life Expectancy
80
60
40
20
0
400
4000
40 000
Income

Let's talk about practical DS and ML

Probably, you have been reading a lot about AI, Big Data, etc., but are you able to apply it on a real research project?

Yes ✅

I'm going to talk about neural networks, deep learning, machine learning, and so on.

Yes, I'm going to talk (less) about statistics and other *classical* approaches, too.

And YES, probably the following contents are opinionated.

Some concepts firts 📚

- What is Data Science?
- What is DL?
- What is ML?
- What is AI?

What is Data Science?

Data science is an inter-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data [1].

Data science is a "concept to unify statistics, data analysis, machine learning and their related methods" to "understand and analyze actual phenomena" with data [2].

What is Deep Learning?

Deep learning (also known as deep structured learning or differential programming) is part of a broader family of machine learning methods based on artificial neural networks with representation learning. Learning can be supervised, semi-supervised or unsupervised [3-5].

What is Machine Learning?

Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead.

It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training data", to make predictions or decisions without being explicitly programmed to perform the task [6, 7].

What is Artificial Intelligence?

"the designing and building of intelligent agents that receive percepts from the environment and take actions that affect that environment" [8].

Major goals of AI

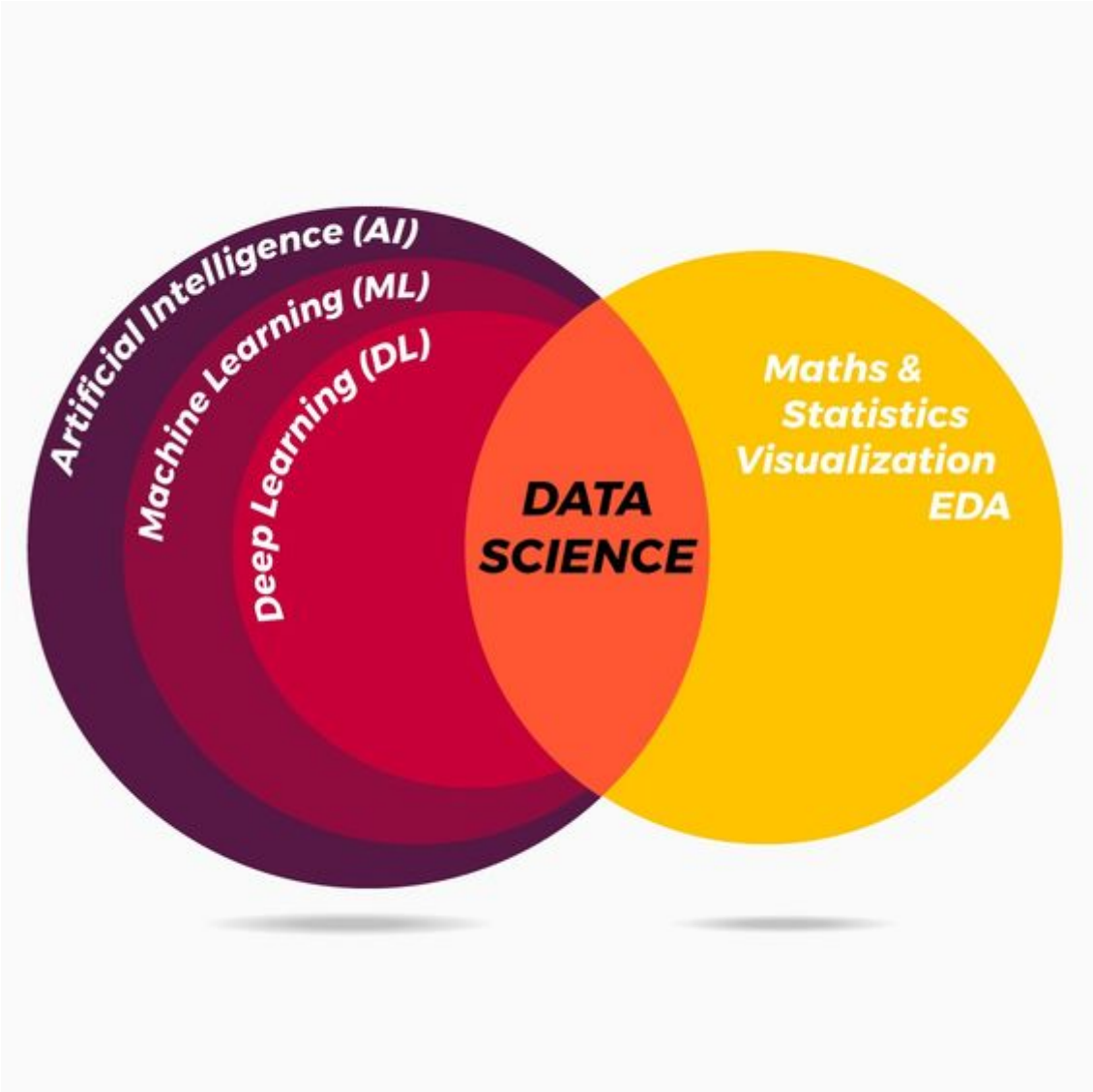
- Knowledge reasoning
- Planning
- Machine learning
- Natural language processing
- Computer vision
- Robotics
- Artificial general intelligence

Common points of view (mostly wrong)

Most people assume that statistics and machine learning are only related to fitting (math) models.

Most (illiterate) people assume that AI === neural networks

Relationship between all of those concepts



Source. https://www.corpnce.com/category/artificial-intelligence/ (https://www.corpnce.com/category/artificial-intelligence/)

Ok, and now what? 🤷‍♀️🤷‍♂️

Let's talk about methods 👩‍🏫 , tools ⚒️ , resources ✨ , courses 🎓 , examples 🏗️ , etc.

To do so, we're going to focus on some areas

- Classification & Clustering
- Computer Vision
- Natural Language Processing
- Reinforcement Learning

Classification & Clustering 🎯

Classification & Clustering 🎯

In machine learning and statistics, classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known [9]

Classification is considered an instance of supervised learning, i.e., learning where a training set of correctly identified observations is available. The corresponding unsupervised procedure is known as **clustering**, and involves grouping data into categories based on some measure of inherent similarity or distance [9].

Classification. Methods & algorithms 🧑‍🔬

- Linear classifiers
    - Fisher's linear discriminant
    - logistic regression
    - Naive Bayes classifier
    - Perceptron
- Support vector machines
    - Least squares support vector machines
- Quadratic classifiers

Classification. Methods & algorithms 🧑‍🔬

- Kernel estimation
    - k-nearest neighbor
- Boosting (meta-algorithm)
- Decision trees
    - Random forests
- Neural networks
- Learning vector quantization

Clustering. Methods & algorithms 👩‍🔬

- Centroid-based (K-means)
- Connectivity-based (hierarchical clustering)
- Distribution-based clustering
- Density-based clustering

Classification & Clustering. Methods & algorithms 👩‍🔬

My favourite options:

- Random forests (classification)
- XGBoost (classification)
- Hierarchical clustering (clustering)

Classification & Clustering. Recommended tools ⚒️

For almost anything related to classification, clustering, regressions, etc.: **Scikit-learn**
https://scikit-learn.org/stable/ (https://scikit-learn.org/stable/)

For XGBoost and Gradient Boosting-related algorithms, the original XGBoost package:
https://xgboost.readthedocs.io/en/latest/ (https://xgboost.readthedocs.io/en/latest/)

Classification & Clustering. Resources ✨

- Scikit-learn user guide: https://scikit-learn.org/stable/user_guide.html (https://scikit-learn.org/stable/user_guide.html)
- Kaggle course on Intro to Machine Learning: https://www.kaggle.com/learn/intro-to-machine-learning (https://www.kaggle.com/learn/intro-to-machine-learning)
- Fastai's course: Introduction to Machine Learning for Coders! https://course18.fast.ai/ml.html (https://course18.fast.ai/ml.html)

Classification & Clustering. Examples 🏗️

- Random Forests & Hierarchical Clustering: [github.com/cbjuan/paper-ieeeAccess-2017 (https://github.com/cbjuan/paper-ieeeAccess-2017/blob/master/machinelearning-results.ipynb)](https://github.com/cbjuan/paper-ieeeAccess-2017/blob/master/machinelearning-results.ipynb)
- XGBoost: [github.com/datalabusal2018/MachineLearningTest/ (https://github.com/datalabusal2018/MachineLearningTest/blob/juancb/DataLab%20ML%20-%20XGBoost.ipynb)](https://github.com/datalabusal2018/MachineLearningTest/blob/juancb/DataLab%20ML%20-%20XGBoost.ipynb)

Computer Vision 👀

Computer Vision 👀

Computer vision is an interdisciplinary scientific field that deals with how computers can be made to gain high-level understanding from digital images or videos. From the perspective of engineering, it seeks to automate tasks that the human visual system can do [10-12].

Computer Vision. Typical tasks ✔

- Recognition
- Motion analysis
- Scene reconstruction
- Image restoration

Computer vision. Methods & algorithms 👷

- Recognition
    - Object recognition (object classification)
    - Identification
    - Detection

  Current best approaches:

    - Convolutional neural networks (or other specialized NNs)
    - OCR

Computer vision. Methods & algorithms 👩‍🔬

- Motion analysis (motion estimation)
- Scene reconstruction (from images or videos, build 3D environments)
- Image restoration

Current research areas: applying deep learning to them

Classification & Clustering. Recommended tools ⚒

To get into all of those things, especially in detection, I recommend FastAI https://www.fast.ai/ (https://www.fast.ai/)

If you want to play with image restoration, check out DeOldify https://github.com/jantic/DeOldify (https://github.com/jantic/DeOldify)

If you want to go deeper, build your own NNs using FastAI, PyTorch (https://pytorch.org/ (https://pytorch.org/)) or Tensorflow (https://www.tensorflow.org/ (https://www.tensorflow.org/)) / Keras (https://keras.io/ (https://keras.io/))

Computer Vision. Resources ✨

- FastAI course on Practical Deep Learning for Coders https://course.fast.ai/ (https://course.fast.ai/)
- PyTorch tutorial for beginners: Transfer Learning for Computer Vision Tutorial https://pytorch.org/tutorials/beginner/transfer_learning_tutorial.html (https://pytorch.org/tutorials/beginner/transfer_learning_tutorial.html)

Computer Vision. Examples 🏗️

- Image classification (pets): https://github.com/fastai/course-v3/blob/master/nbs/dl1/lesson1-pets.ipynb (https://github.com/fastai/course-v3/blob/master/nbs/dl1/lesson1-pets.ipynb)
- Head pose estimation via regression: https://github.com/fastai/course-v3/blob/master/nbs/dl1/lesson3-head-pose.ipynb (https://github.com/fastai/course-v3/blob/master/nbs/dl1/lesson3-head-pose.ipynb)

Natural Language Processing 📖

Natural Language Processing 📖

NLP is a principled approach to processing human languages. Formally, it is a subfield of Artificial Intelligence (AI) that refers to computational approaches to process, understand, and generate human languages. It is a subfield of AI because processing language is considered to be a huge part of human intelligence. Use of language is arguably the most salient skill that separates humans from other animals [13].

NLP. Tasks and areas ✓

- Syntax
    - Lemmatization
    - Morphological segmentation
    - Part-of-speech tagging
    - Parsing
    - Sentence breaking (boundary disambiguation)
    - Stemming
    - Terminology extraction
    - ...

NLP. Tasks and areas ✓

- Semantics
    - Machine translation
    - Natural language generation / understanding
    - Question answering
    - Relationship extraction
    - Sentiment analysis
    - Topic segmentation and recognition
    - Word sense disambiguation
    - ...

NLP. Tasks and areas ✓

- Discourse
    - Automatic summarization
    - Discourse analysis
- Speech
    - Speech recognition
    - Speech segmentation
    - Text-to-speech
- Dialogue

NLP. Methods & algorithms 👩‍🔬

- Rule based (heuristics)
- Statistical NLP
    - Deep Learning-based NLP

NLP. Recommended tools ⚒️

To get into all of those things, I recommend SpaCy https://spacy.io/ (https://spacy.io/)

Many other big libraries: StandfordNLP (https://stanfordnlp.github.io/stanfordnlp/ (https://stanfordnlp.github.io/stanfordnlp/)), NLTK (https://www.nltk.org/ (https://www.nltk.org/)), AllenAI NLP (https://allennlp.org/ (https://allennlp.org/)), FastAI, HuggingFace Transformers (https://github.com/huggingface/transformers (https://github.com/huggingface/transformers)), etc.

NLP. Examples 🏗️

- Get topics via LDA. Using SpaCy: https://github.com/felicidadgsanchez/visual-literacy-survey-2018/blob/master/notebooks/prosume_all.ipynb (https://github.com/felicidadgsanchez/visual-literacy-survey-2018/blob/master/notebooks/prosume_all.ipynb)
- Language generation via FastAI: https://github.com/fastai/course-v3/blob/master/nbs/dl1/lesson3-imdb.ipynb (https://github.com/fastai/course-v3/blob/master/nbs/dl1/lesson3-imdb.ipynb)
- HuggingFace. Write with Transformers: https://transformer.huggingface.co/ (https://transformer.huggingface.co/)

Reinforcement Learning 🎮

Reinforcement Learning 🎮

Reinforcement learning (RL) is the problem faced by an agent that must learn behavior through trial-and-error interactions with a dynamic environment [14].

Reinforcement learning is one of three basic machine learning paradigms, alongside supervised learning and unsupervised learning.

Reinforcement Learning. Methods & algorithms 👩‍🏭

- Monte Carlo
- Q-learning
- SARSA
- DQN
- DDPG
- A3C
- Others based on Deep Learning, etc.

Reinforcement Learning. Recommended tools ⚒

- OpenAI Gym https://gym.openai.com/ (https://gym.openai.com/)
- Unity ML Agents https://github.com/Unity-Technologies/ml-agents (https://github.com/Unity-Technologies/ml-agents)

Reinforcement Learning. Examples 🏗️

- DeepMind Alpha Go https://deepmind.com/research/case-studies/alphago-the-story-so-far (https://deepmind.com/research/case-studies/alphago-the-story-so-far)
- Unity ML Agents https://unity3d.com/machine-learning/ (https://unity3d.com/machine-learning/)
- Andrej Karpathy. Deep Reinforcement Learning: Pong from Pixels https://karpathy.github.io/2016/05/31/rl/ (https://karpathy.github.io/2016/05/31/rl/)

General resources and advices

- Take **a lot of care** about data cleaning and preparation. In 🐍 -> 🐼
- Practice EDA (Exploratory Data Analysis). In 🐍 -> 🐼
- Visualize your data and models
    - Pandas, Seaborn, Plotly, etc.
- Visualize your trainings
    - Tensorboard
    - FastAI
    - Comet ML
    - Neptune.ai

General resources and advices

- Prepare properly datasets (training, validation, test)
- Take care about validation against benchmarks, etc.
- Consider the resources needed for training & developing 💵 ?
- Consider the resources needed for deploying your solutions/models 💵 ?

Conclusions & thoughts

There is **a lot to do** in DS, ML, DL, AI, etc. They are hot topics, and they will be for a time.

There are different areas of research depending on whether you want to do theoretical research or applied research

There are a lot of interesting problems to solve. You should find out what motivates you

References:

[1] Dhar, V. (2013). Data science and prediction. Communications of the ACM, 56(12), 64-73.

[2] Hayashi, C. (1998). What is data science? Fundamental concepts and a heuristic example. In Data science, classification, and related methods (pp. 40-51). Springer, Tokyo.

[3] Bengio, Y.; Courville, A.; Vincent, P. (2013). "Representation Learning: A Review and New Perspectives". IEEE Transactions on Pattern Analysis and Machine Intelligence. 35 (8): 1798–1828. arXiv:1206.5538. doi:10.1109/tpami.2013.50.

References:

[4] Schmidhuber, J. (2015). "Deep Learning in Neural Networks: An Overview". Neural Networks. 61: 85–117. arXiv:1404.7828. doi:10.1016/j.neunet.2014.09.003.

[5] Bengio, Yoshua; LeCun, Yann; Hinton, Geoffrey (2015). "Deep Learning". Nature. 521 (7553): 436–444. Bibcode:2015Natur.521..436L. doi:10.1038/nature14539.

[6] Koza, John R.; Bennett, Forrest H.; Andre, David; Keane, Martin A. (1996). Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming. Artificial Intelligence in Design '96. Springer, Dordrecht. pp. 151–170.

References

[7] Bishop, C. M. (2006), Pattern Recognition and Machine Learning, Springer, ISBN 978-0-387-31073-2

[8] Russell, S. J., & Norvig, P. (2016). Artificial intelligence: a modern approach. Malaysia; Pearson Education Limited.

[9] Alpaydin, Ethem (2010). Introduction to Machine Learning. MIT Press. p. 9. ISBN 978-0-262-01243-0.

References

[10] Dana H. Ballard; Christopher M. Brown (1982). Computer Vision. Prentice Hall. ISBN 978-0-13-165316-0.

[11] Huang, T. (1996-11-19). Vandoni, Carlo, E (ed.). Computer Vision : Evolution And Promise (PDF). 19th CERN School of Computing. Geneva: CERN. pp. 21–25. doi:10.5170/CERN-1996-008.21. ISBN 978-9290830955.

[12] Milan Sonka; Vaclav Hlavac; Roger Boyle (2008). Image Processing, Analysis, and Machine Vision. Thomson. ISBN 978-0-495-08252-1.

[13] Masato Hagiwara (2020). Real-World Natural Language Processing. https://www.manning.com/books/real-world-natural-language-processing (https://www.manning.com/books/real-world-natural-language-processing)

References

[14] Kaelbling, Leslie P.; Littman, Michael L.; Moore, Andrew W. (1996). "Reinforcement Learning: A Survey". Journal of Artificial Intelligence Research. 4: 237–285. arXiv:cs/9605103.

Links

- Machine-Learning-Tokyo / AI_Curriculum - https://github.com/Machine-Learning-Tokyo/AI_Curriculum (https://github.com/Machine-Learning-Tokyo/AI_Curriculum)
- HuggingFace. Write with Transformers. https://transformer.huggingface.co/ (https://transformer.huggingface.co/)
- An opinionated guide to ML Research http://joschu.net/blog/opinionated-guide-ml-research.html (http://joschu.net/blog/opinionated-guide-ml-research.html)

Links

- How to build SOTA conversational AI with transfer learning https://medium.com/huggingface/how-to-build-a-state-of-the-art-conversational-ai-with-transfer-learning-2d818ac26313 (https://medium.com/huggingface/how-to-build-a-state-of-the-art-conversational-ai-with-transfer-learning-2d818ac26313)
- HuggingFace Transformers https://github.com/huggingface/transformers (https://github.com/huggingface/transformers)
- Jay Alammar's blog https://jalammar.github.io/ (https://jalammar.github.io/)
- Seb Ruder. 10 ML & NLP Research Highlights of 2019 https://ruder.io/research-highlights-2019/ (https://ruder.io/research-highlights-2019/)

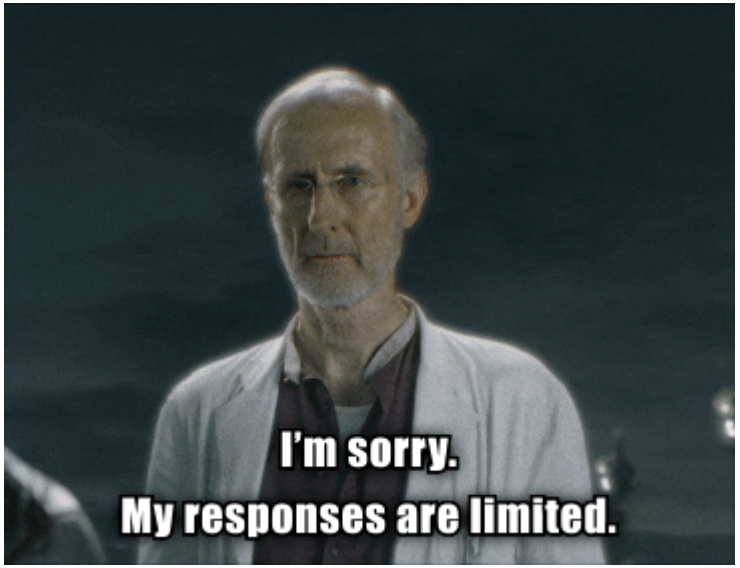
Links

- Sebastian Raschka Deep Learning Models https://github.com/rasbt/deeplearning-models (https://github.com/rasbt/deeplearning-models)
- CS224N: Natural Language Processing with Deep Learning @ Stanford NLP https://www.youtube.com/playlist?list=PLoROMvodv4rOhcuXMZkNm7j3fVwBBY42z&utm_campaign=NLP+News&utm_medium=email&utm_source=Revue+newsletter (https://www.youtube.com/playlist?list=PLoROMvodv4rOhcuXMZkNm7j3fVwBBY42z&utm_campaign=NLP+News&utm_medium=email&utm_source=Revue+newsletter)

Link to presentation and materials

https://github.com/cbjuan/talk-ai-mis-usal-2020 (https://github.com/cbjuan/talk-ai-mis-usal-2020)

Questions?



🤖 Practical Data Science and Machine Learning 🤖

Juan Cruz-Benito

🔭 @_juancb

Master on Intelligent Systems. University of Salamanca. 27 Feb 2020