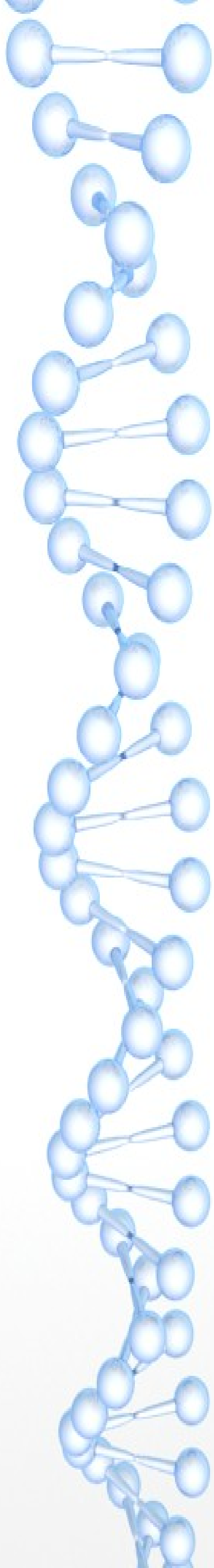


A **very short** introduction to Machine Learning And Artificial Intelligence and, by the way, Data Science

José R. Valverde

CNB-CSIC
Feb 2024

Relax

Image by pendleburyannette from Pixabay

There is no reason to fret

What this is all about

- Often both terms (ML/AI) are used interchangeably

  (you have been warned!)
- We start from *Big data* (large/huge datasets)
- *Data mining* explores and searches for useful associations in big data.
- *Machine learning* is a subset of AI that allows machines to learn from data without human help.
- *Artificial Intelligence* is a broader concept, includes ML but also human-assisted learning and all processes that make machines look “intelligent” (spoiler: they aren’t).
- *Data Science* tries to make sense of data (find out associations, build models, use them..).

Learning

- Supervised
  - A human trains the computer
  - More efficient and powerful
  - Cumbersome to set up
- Unsupervised
  - The computer “learns” by itself
  - Less efficient and powerful
  - Does not require human supervision

Unsupervised learning

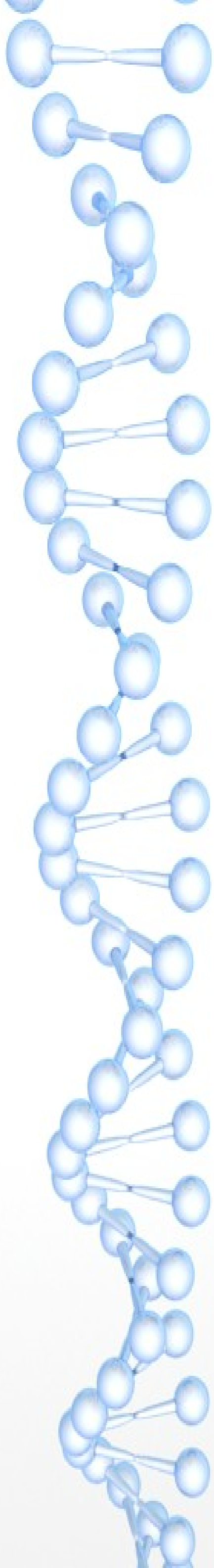
- We use the machine to carry out complex “intelligence” tasks, such as…
  - Identifying relevant variables
  - Identifying variable importance for an outcome
  - Identifying associations and relations
  - Classifying and grouping data, etc…
- Sounds familiar?
  - Regression, statistical models, clustering, factor analysis, PCA, CCA, DeepCCA, EFA, etc...

Caution! Here be dragons!

Image by pendleburyannette from Pixabay

Blind!

- Unsupervised learning does not require any human intervention
- Computers do not “understand” what they do
- Only associations are considered, **not causality**
- There is no underlying reasoning.
- Choices may be wrong by pure **chance**
- At the end you should look at the associations yourself and decide if they do make sound scientific sense
- Hint: it is all a matter of **interpretation!**



Interpretation

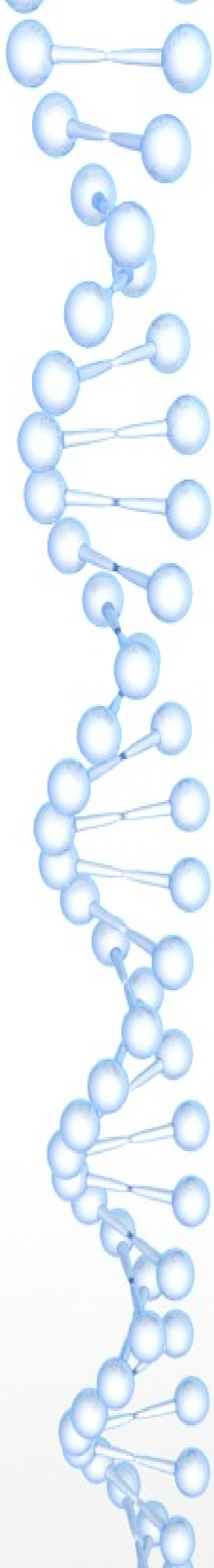
- In unsupervised machine learning, everything depends on how we, humans, **interpret** the associations automatically discovered by the computer.
  - If we fail to interpret correctly the associations discovered by the computer, all the process will be faulty (it does not think for us).

Supervised learning

- Learning data is “labelled” with the solution.
  - Humans must ensure labels are correct before training the computer
- The computer can use these labels to identify which associations are meaningful
- Learning is faster and has less errors
- It’s not that we do not need to, but that often we just cannot supervise the decisions that the computer has done afterwards.

Interpretation again

- The success of supervised learning depends on how we (humans) **interpret** reality before training: we will automate *our* **interpretation** using the computer.
  - If our interpretation of reality is faulty or prejudiced, so will be anything the computer does (it will just reproduce our mistakes).



The decision process

- We want to ensure that we can trust the computer.
  - With unsupervised learning we have to see if the predictions make sense and verify them experimentally
  - With supervised learning we can check how well the computer matches the labels provided to verify it has actually learnt (does reproduce *our pre-hoc* model faithfully).

Artificial Intelligence skeleton

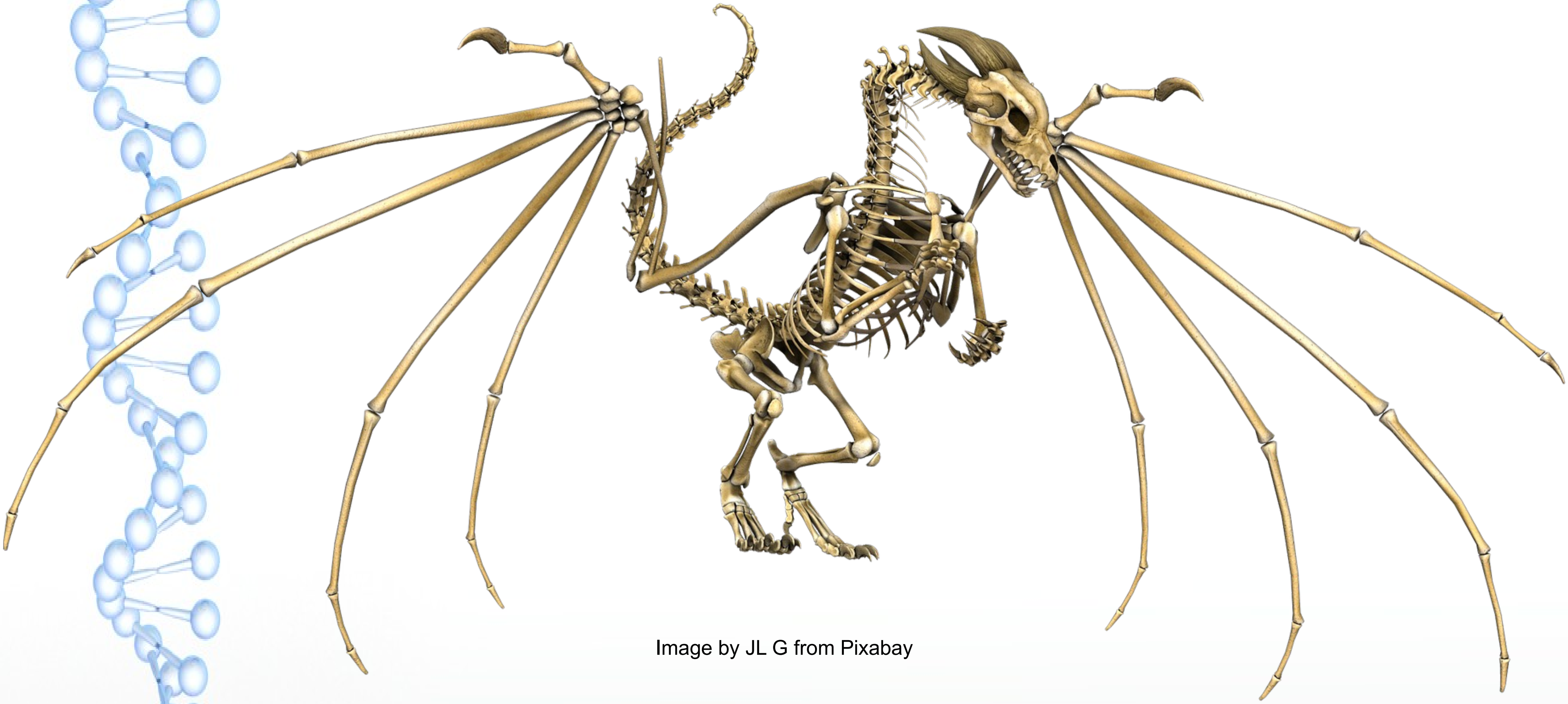


Image by JL G from Pixabay

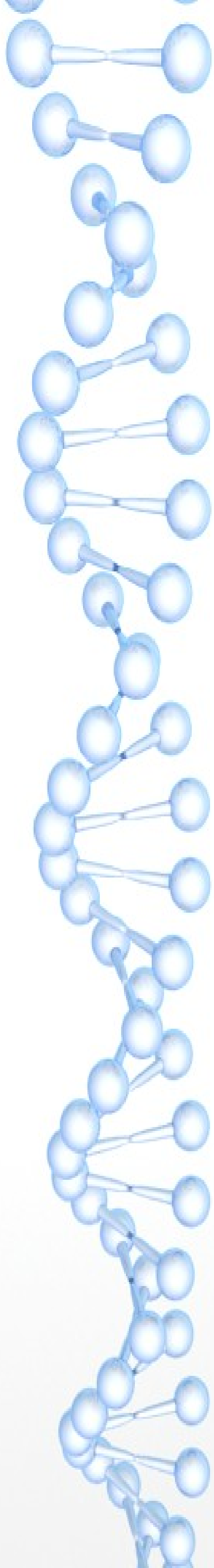
The unsupervised learning process

- Collect the data
- Feed the data to the computer
- Collect the results
- Think about and interpret them

- Easy, isn’t it? Except for the thinking bit!

The supervised training process

- Works like a school:
  - We explain the computer what associations we are interested in:
    - We provide a set of problem data and the correct solutions
    - We repeat it many times so the computer has a chance to learn
  - At key times, we interrupt the process and require the computer to pass a test
    - We give it problem data and hide the answers
    - Then check its answers and grade the exam
  - If it has learned then we can proceed

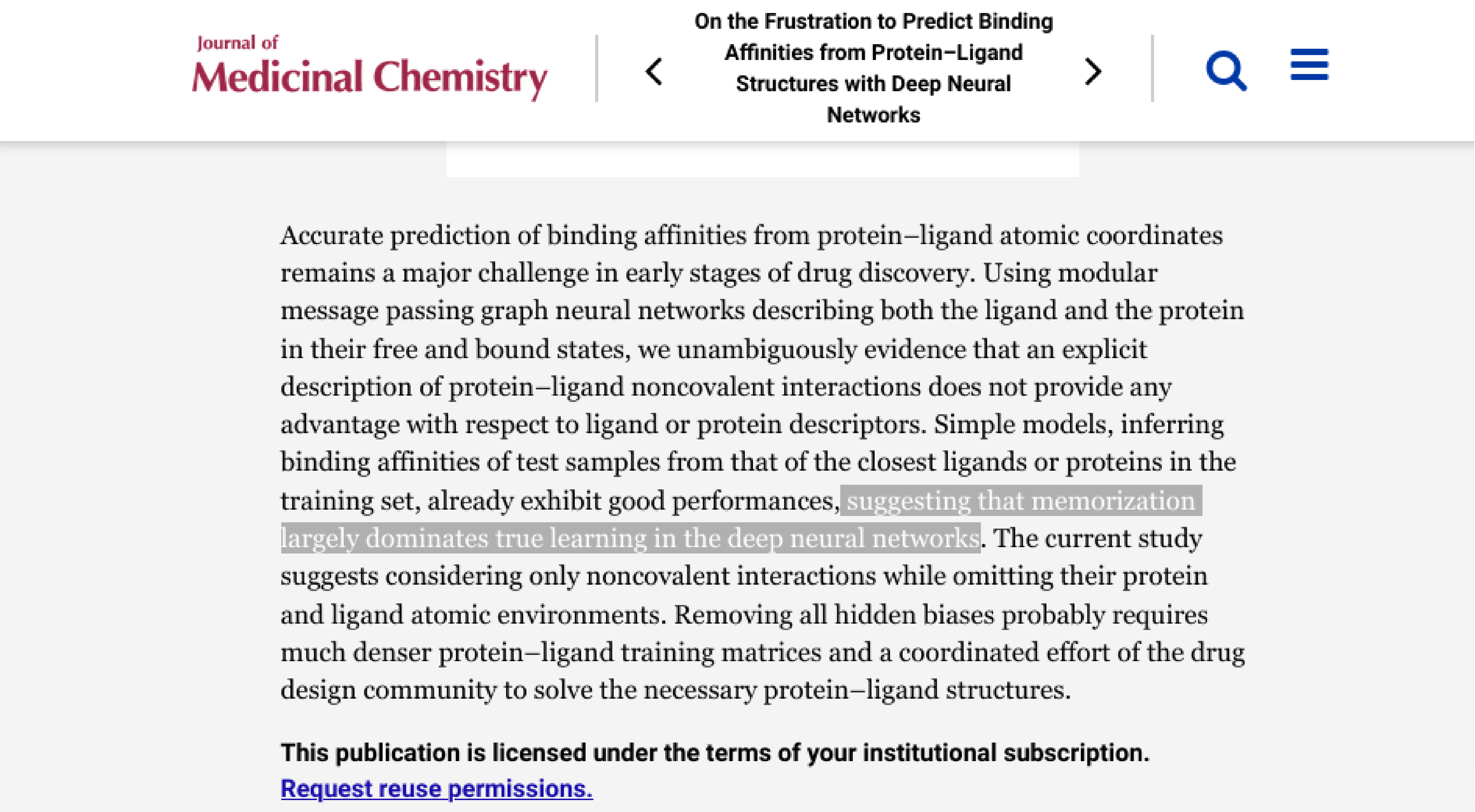


Corollaries

- Before starting supervised learning we need to have a data set large enough with problem data, already correctly solved by trusted humans.
- To avoid cheating we cannot show all the problems and solutions to the computer
  - We typically show ~2/3 or ¾
  - The rest are hidden and set aside for testing
  - During learning we provide problems and answers
  - During testing we provide problems and hide the right answers
  - Afterwards we use the answers withheld to grade the computer performance.

Here be dragons!

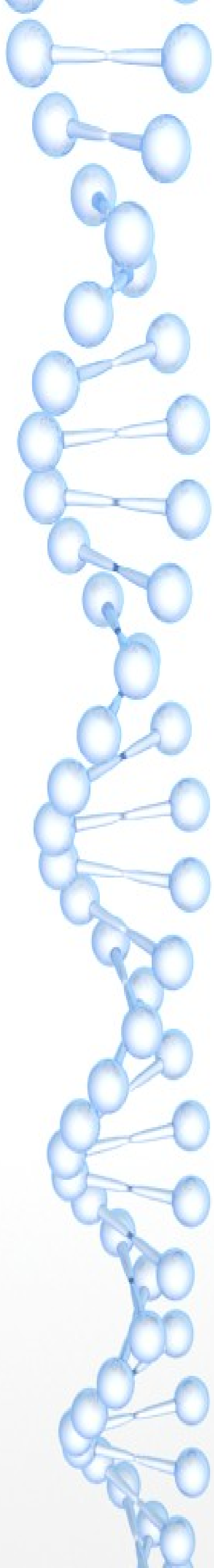
- Modern computers are fast and have a lot of memory
  - Too much
- Sometimes the computer does not “learn”, it just “memorizes” the answers
  - Will be able to answer correctly known questions
  - Will have a higher likelihood of failure with unseen problems



On the Frustration to Predict Binding Affinities from Protein–Ligand Structures with Deep Neural Networks

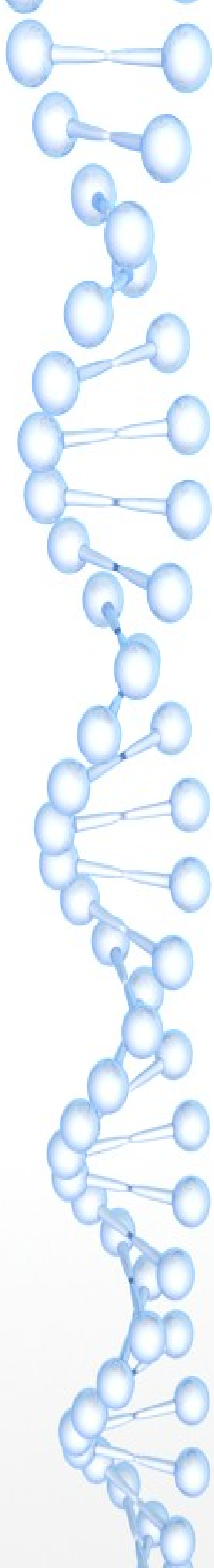
Accurate prediction of binding affinities from protein–ligand atomic coordinates remains a major challenge in early stages of drug discovery. Using modular message passing graph neural networks describing both the ligand and the protein in their free and bound states, we unambiguously evidence that an explicit description of protein–ligand noncovalent interactions does not provide any advantage with respect to ligand or protein descriptors. Simple models, inferring binding affinities of test samples from that of the closest ligands or proteins in the training set, already exhibit good performances, suggesting that memorization largely dominates true learning in the deep neural networks. The current study suggests considering only noncovalent interactions while omitting their protein and ligand atomic environments. Removing all hidden biases probably requires much denser protein–ligand training matrices and a coordinated effort of the drug design community to solve the necessary protein–ligand structures.

J. Med. Chem. 2022, 65, 11, 7946–7958



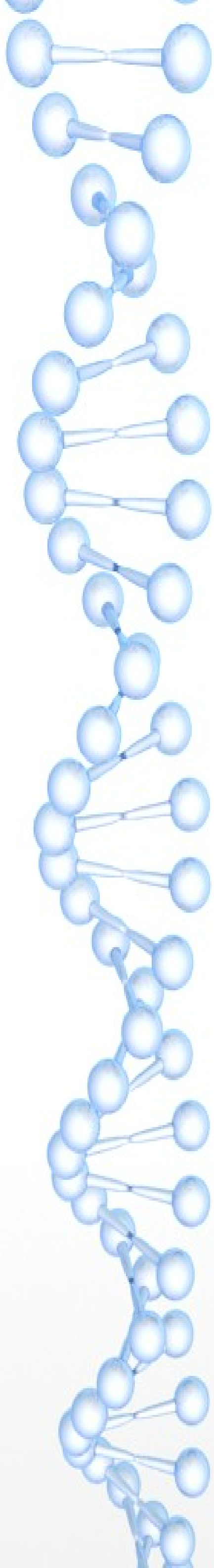
Avoiding over-training

- The problem arises when we
  - over-dimension the learning tool and
  - give too few training problems
- Bottom line: **never trust the computer!**
- How do we deal with uncertainty?
  - **Statistics!**



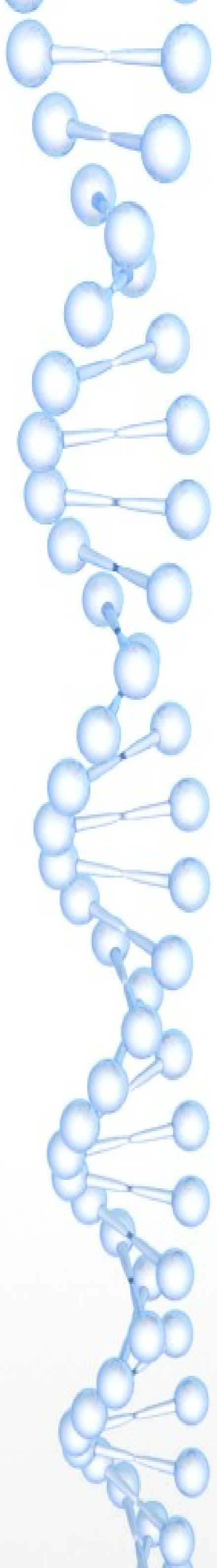
Monitoring over-training

- Repeat the training process many times, randomizing the problems each time
- Each time we will get a different learning model
- Compare them
- Select the best
- Repeat many times
- After grading, use the best one.



Grading

- Take the labeled problems that we set aside at the beginning, hide the answers and give only the problem data to the computer
  - Collect the answers
  - Compare them with the hidden labels/answers
  - Compute statistics
    - Confusion matrix
    - Sensitivity
    - Specificity
    - AUC, ROC, Etc...



The final? test

- Once you have a model that makes satisfactory predictions
  - **Verify them experimentally**
- Once you have experimental proof that it works
  - **Use the system**

The golden rule

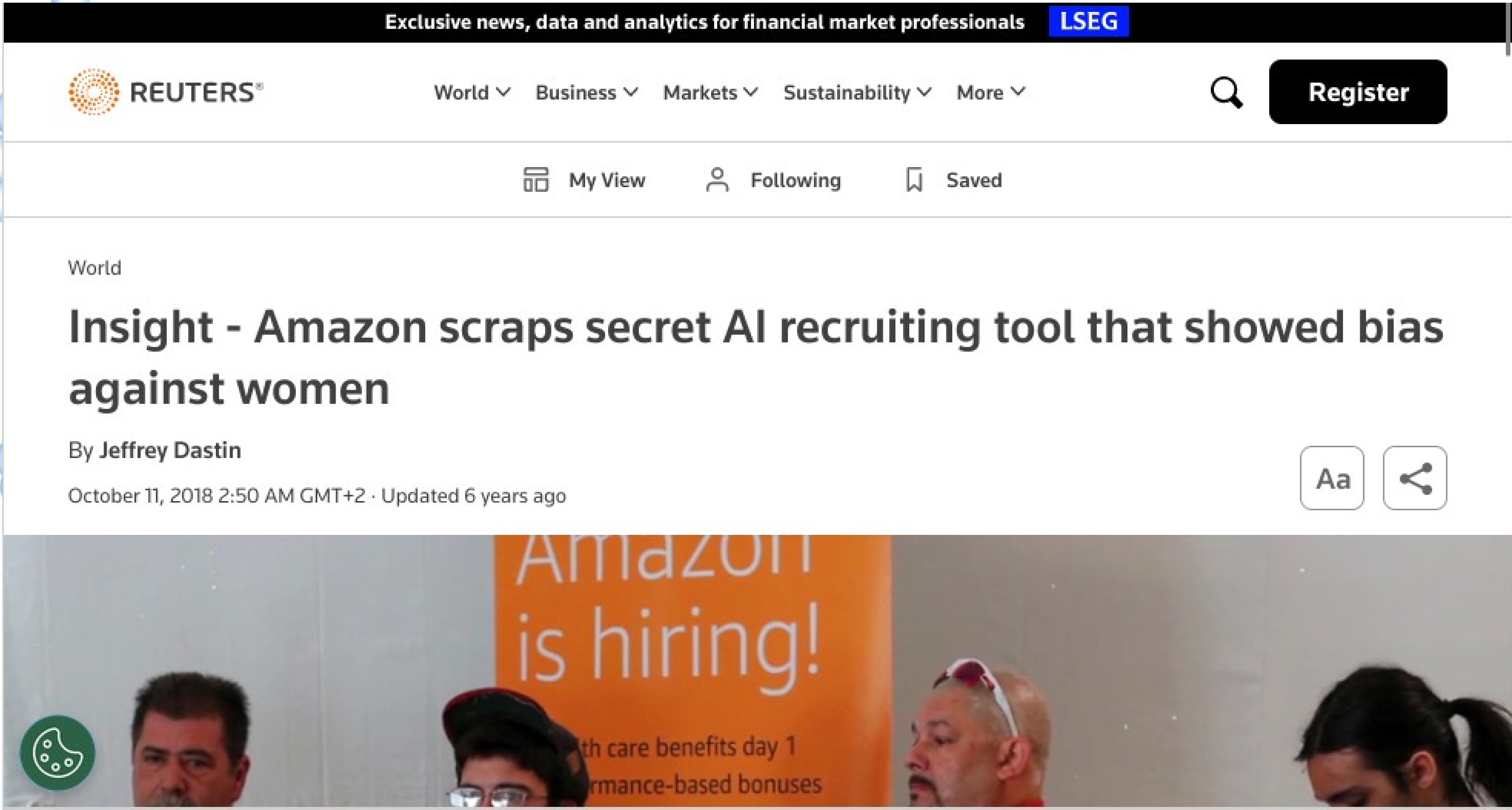
- *There is no sense in Artificial Intelligence when Natural is absent.*

  José R. Valverde

- A model will be only as good as the method used and the knowledge employed to train.
- WE CAN NEVER BE SURE OF ANYTHING WE THINK WE KNOW
  - Neither model, nor data.

Closing the circle

- Remember the Bayesian *motto*?
  - *The posterior of today is the prior of tomorrow*
- *Ergo*, remember to repeat experimental validation of AI models periodically to
  - Revise if the original assumptions used in training are still valid
  - Revise if the answers produced by the model are still valid
    - There is always a non-null chance that the first tests were passed by a random lucky draw of chance.
    - New data may invalidate prior assumptions (but you will only notice if you didn't trust the AI model)

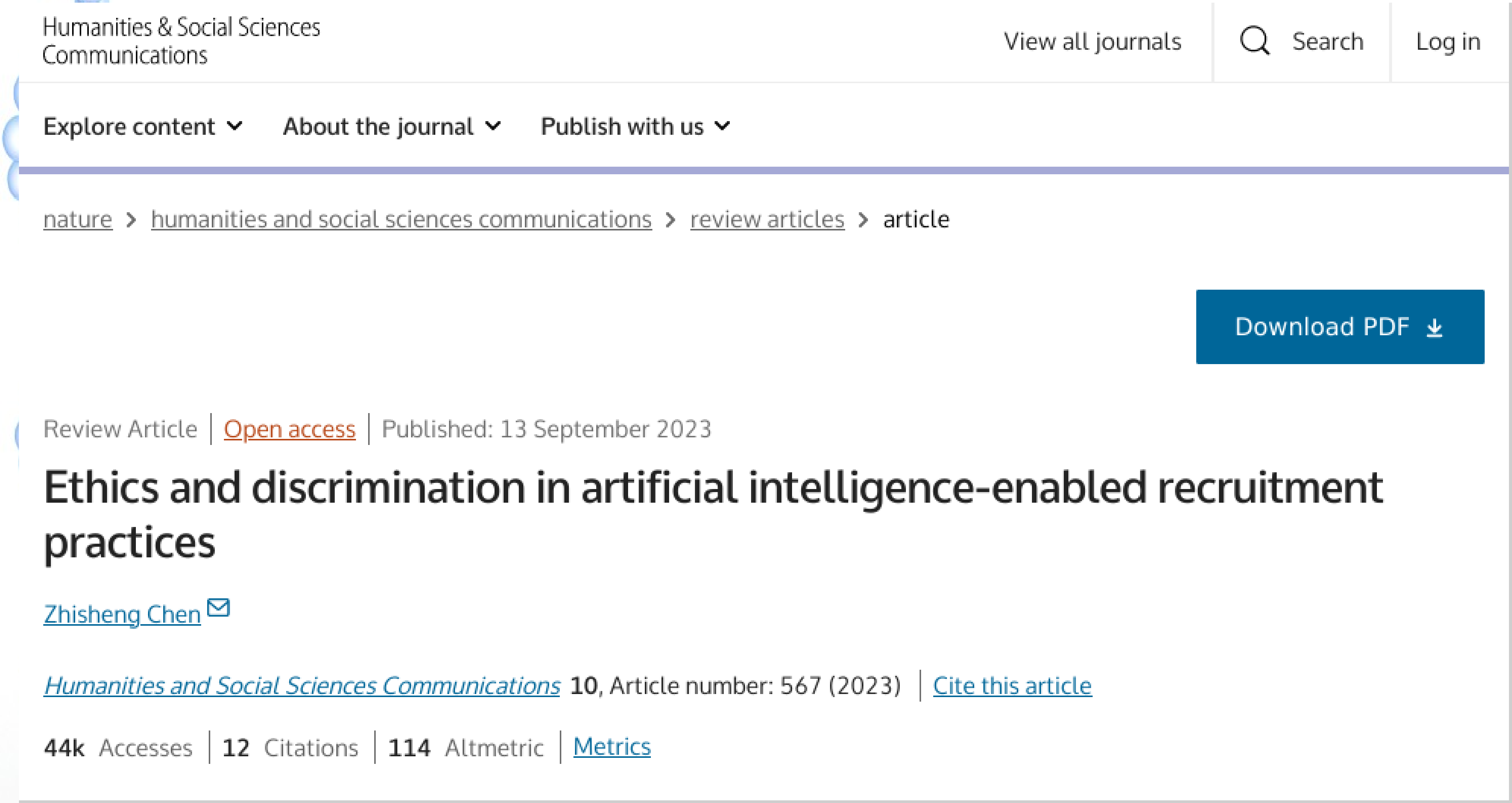


World

Insight - Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin

October 11, 2018 2:50 AM GMT+2 · Updated 6 years ago



Humanities & Social Sciences Communications

nature > humanities and social sciences communications > review articles > article

Review Article | Open access | Published: 13 September 2023

Ethics and discrimination in artificial intelligence-enabled recruitment practices

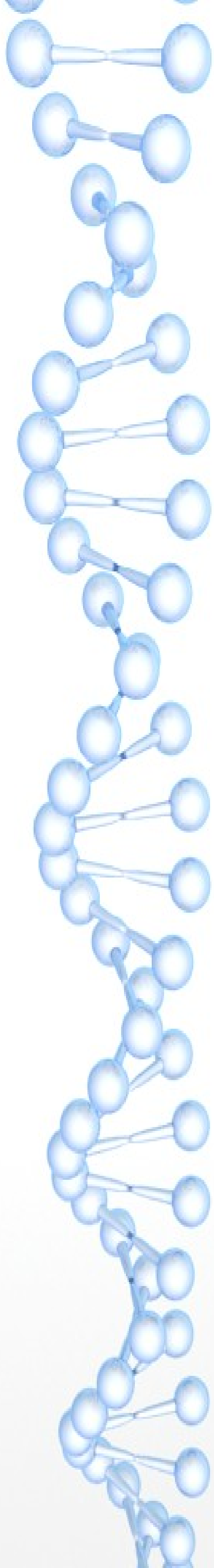
Zhisheng Chen

Humanities and Social Sciences Communications **10**, Article number: 567 (2023) | Cite this article

44k Accesses | 12 Citations | 114 Altmetric | Metrics

Supervised AI summary

- Prepare a well curated large set of training data
- Split it into a training and a testing data set
- Hand over the training data set to the computer and let it learn
  - Repeat the learning process many times
- Hand over the unlabeled test data to the computer and check the answers with the known correct labels.



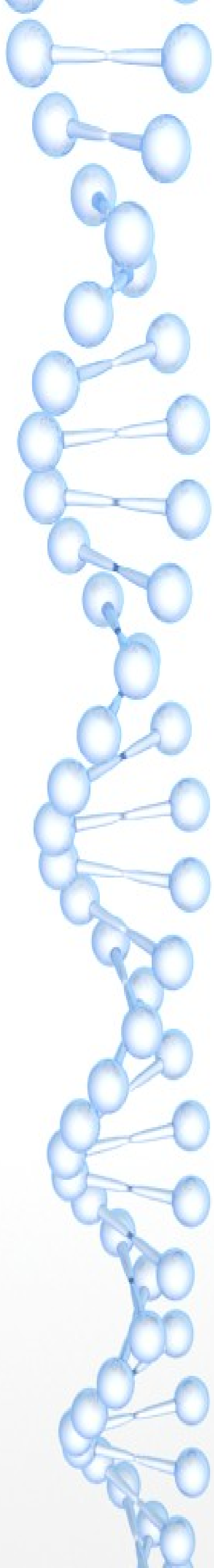
... and remember

Never, ever, trust anyone or anything,
not even yourself, not even the "facts*", much less a computer.

* "facts" (note the quotes) can indeed be misleading or biased: they can be carefully chosen to mislead you (even by you yourself due to unfounded expectations), or your observations be incomplete (limited devices, experience,...). Anybody said "paranoia"? Well, yes! It is your and your descendants' life what is at stake.

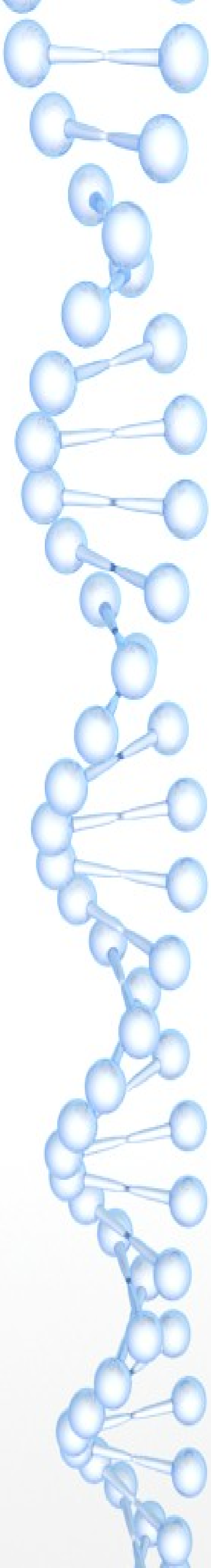
...and certainly, do **not** trust **me!**

- Minority Report?
- https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
- Racist AI?
- https://www.science.org/doi/10.1126/science.aax2342
- Biased selection?
- https://www.bbc.com/news/education-53787203
- Accurate predictions?
- https://www.science.org/doi/10.1126/science.1248506



Practical decisions

- There are literally thousand of learning methods.
- Which one should I use?
  - The right answer: study and understand them, study and understand your problem and make an educated decision
  - The pragmatic answer: read the literature about your problem, look for comparative reviews and use the method reported to be best according to experience.



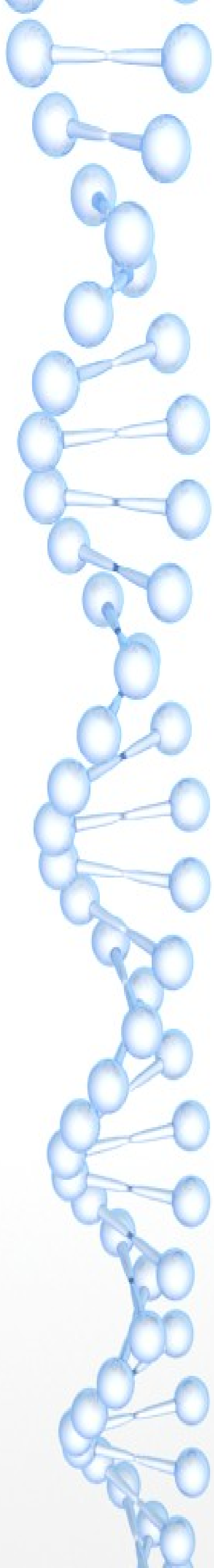
Discovering associations

Well, we already know something

- Correlation/regression
- Multiple regression models
  - Linear, nonlinear, MARS, etc...
- PCA
- EFA
- CCA
- Neural Networks
- etc...

Finding structure

- Cluster analysis
  - Group individuals according to their scoring on different variables
  - Various methods are very popular:
    - Hierarchical clustering
    - K-means
    - Fuzzy C-means
    - Neighborhood based
    - Random Forest
    - DB-cluster
    - Neural Networks
    - etc...

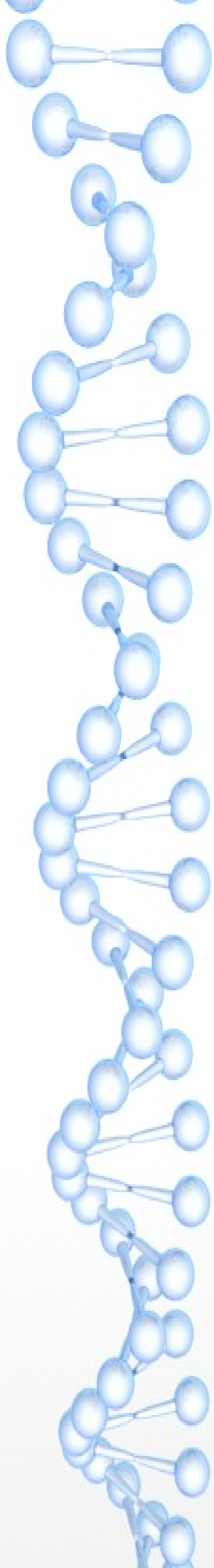


Telling groups apart

- We want to discriminate subjects and separate (classify) them into groups
  - K-Nearest Neighbors
  - Linear Discriminant analysis
  - Decision Tree
  - Random Forest
  - Neural Networks
  - Support Vector Machines (SVMs)

Regression

- Measure degrees of association
  - K-Nearest Neighbors
  - Neural Network
  - Support Vector Machine (SVM)
  - Decision tree
  - Random Forest
  - Regularized linear regression
  - MARS (EARTH)
  - Neural Networks
  - ...



Prediction/forecasting

- Usually used as an extension of longitudinal analyses/time series
  - Linear models
  - Non-linear models
  - ARIMA models
  - MARS (EARTH) models
  - Neural Networks
  - K-nearest neighbors
  - -etc...

Translating "languages"

- Translate any language (not just human, e.g. the "languages of life")
- LLM (Large Language Models)
  - AlphaFold and derivatives
  - LLαMA 2
  - Falcon
  - Flor
  - BERT
  - GPT, GPT-NeoX, GPT-J etc...

Questions?