Machine learning, Galaxy and more

Anup Kumar

Bioinformatics group, University of Freiburg, Freiburg, Germany

26th October 2021

Agenda

- Basics of machine learning
- Machine learning in Galaxy
- Ongoing machine learning projects - Jupyterlab editor for ML and predicting protein evolution in SARS-COV2 sequences using deep learning

Basics of Machine learning

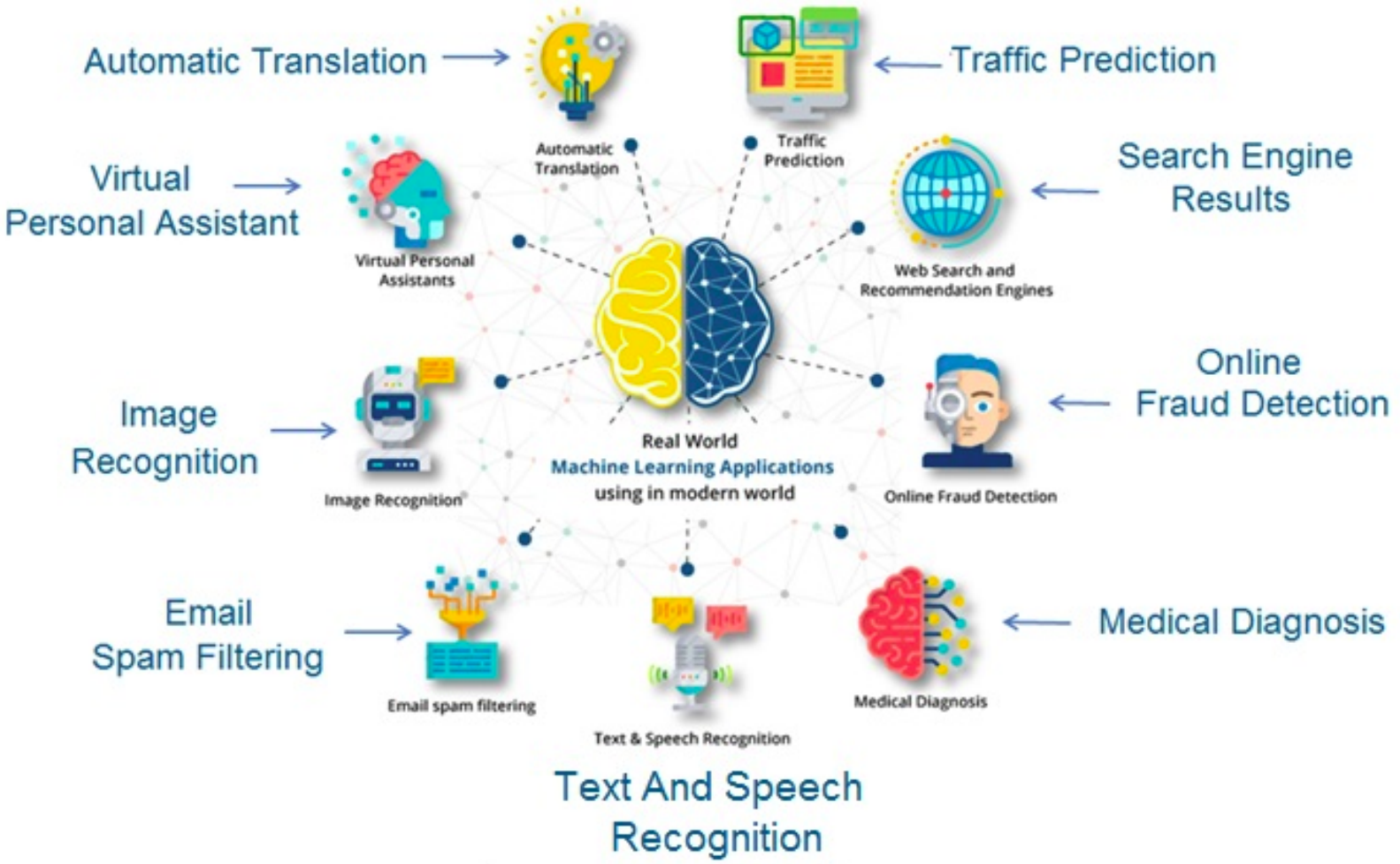
Machine learning (ML)

- ML - computer program that learns rules from data
- Use rules to distinguish patterns
- Rules are mathematical functions
- Learn on existing (training) data, predict unknown outputs (of test data)
- ML algorithms work on numbers and not text or characters
- Example task: handwritten digit recognition

1

[1] https://en.wikipedia.org/wiki/MNIST_database

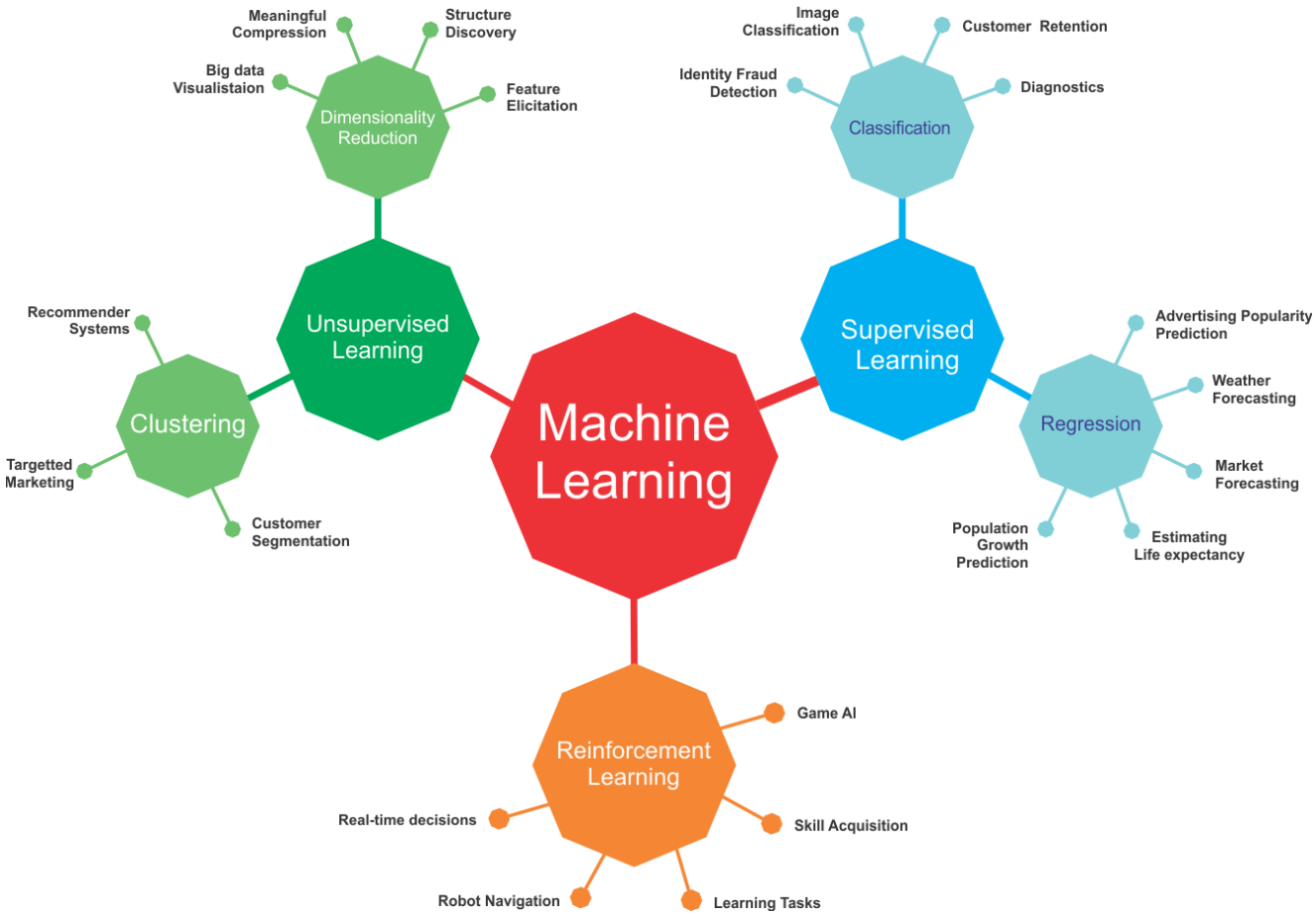
General applications of ML



2

[2] https://www.learncomputerscienceonline.com/what-is-machine-learning/

Types of ML



3

[3] https://skilllx.com/list-of-machine-learning-algorithms/

Supervised learning (Classification)

- Labeled data
- Features (gender, height, weight, index)
- Labels/classes/targets/output (status)

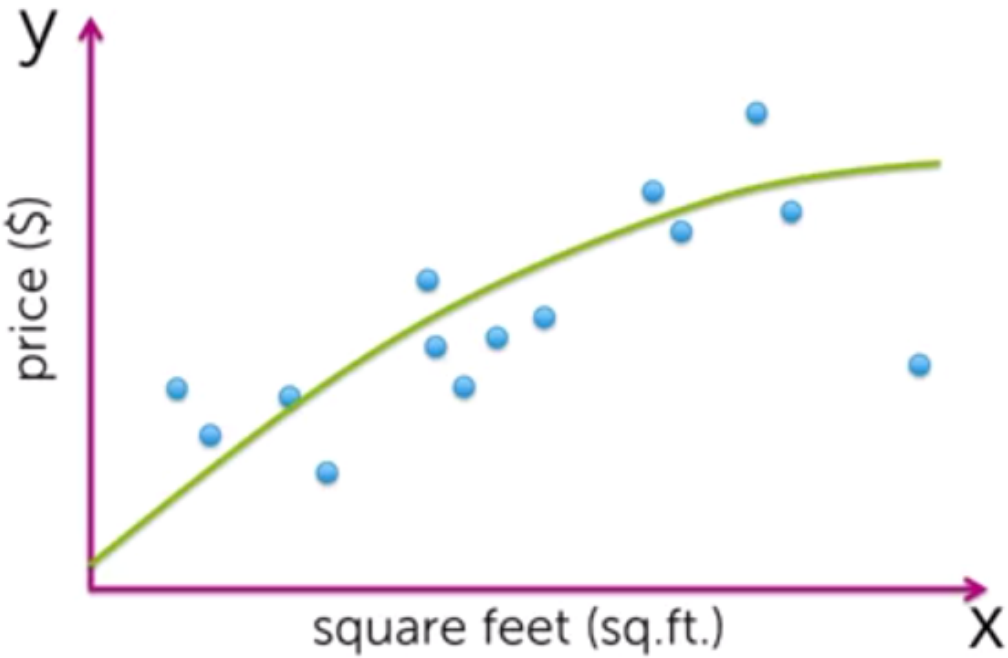
	Gender	Height	Weight	Index	Status
0	Male	174	96	4	Obesity
1	Male	189	87	2	Normal
2	Female	185	110	4	Obesity
3	Female	195	104	3	Overweight
4	Male	149	61	3	Overweight
5	Male	189	104	3	Overweight
6	Male	147	92	5	Extreme Obesity
7	Male	154	111	5	Extreme Obesity
8	Male	174	90	3	Overweight
9	Female	169	103	4	Obesity

4

[4] https://www.kaggle.com/yersever/500-person-gender-height-weight-bodymassindex

Supervised learning (Regression)

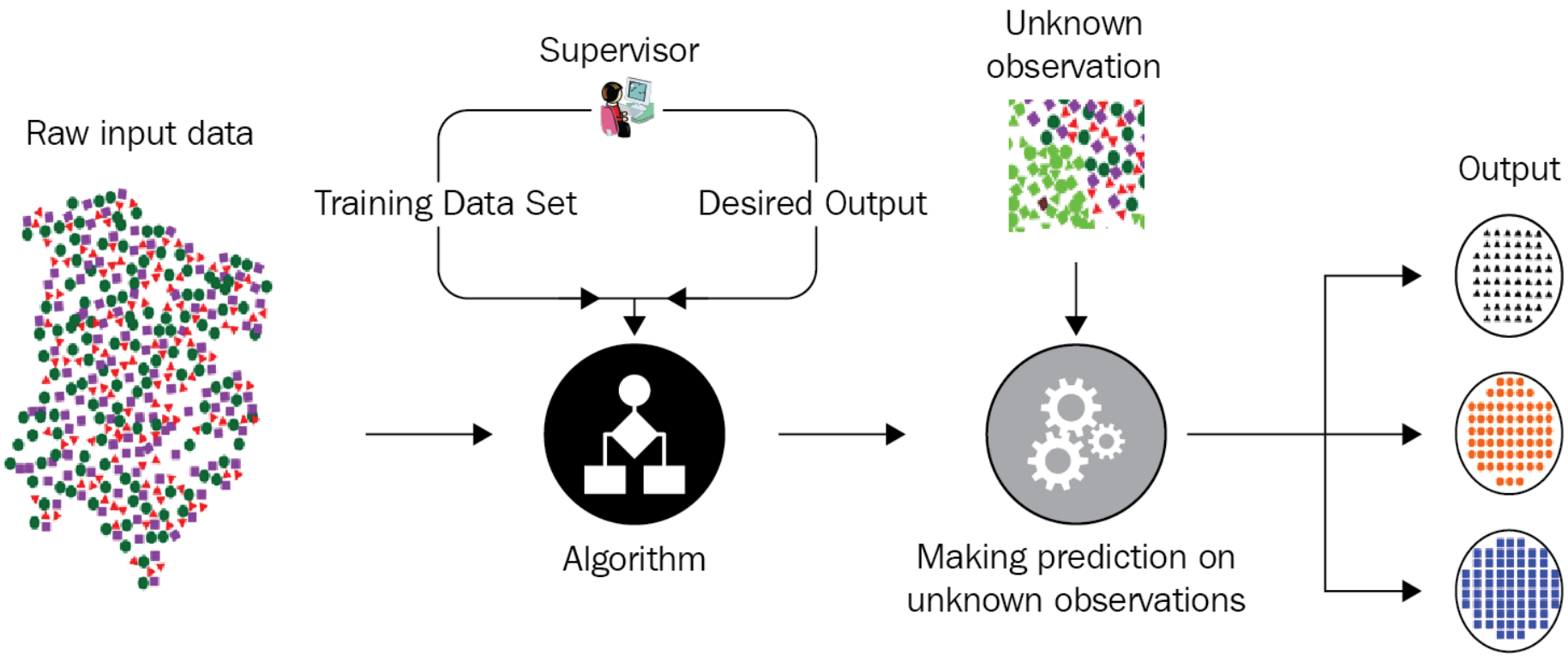
- Labeled data
- Class is a real number instead of a category
- Example: house price prediction



5

[5] https://finalagito.github.io/2016/08/17/The-Linar-Regression-The-First-Step-Of-Machine-Learning/

Pipeline for supervised learning



6

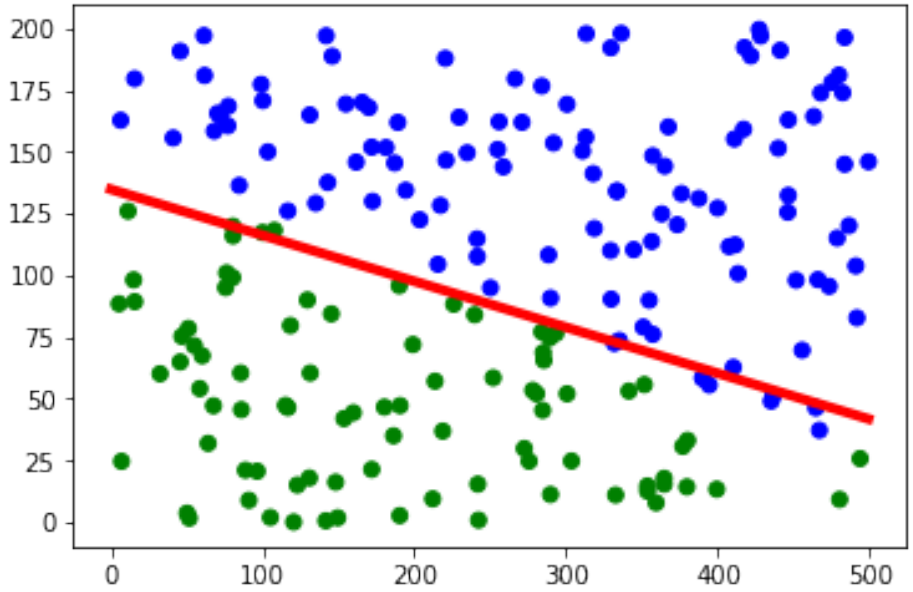
[6] https://subscription.packtpub.com/book/big_data_and_business_intelligence/9781789345070/1/ch01lvl1sec12/ml-tasks

Algorithms for supervised learning

- Linear models
- Support vector machines
- Decision trees
- Ensemble models
- Neural networks
- ...

Linear models

- Learn straight line decision boundary
- Easy to use and fast
- Don't learn non-linear features

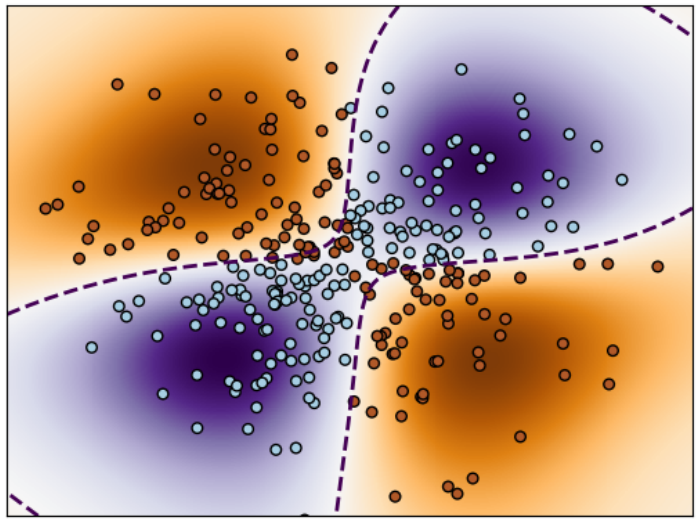


7

[7] https://stats.stackexchange.com/questions/436827/why-does-linear-non-logistic-regression-work-as-a-linear-classifier-what-classi

Non-linear models

- Many times, patterns can only be separated by non-linear boundaries
- Linear models are not sufficient
- Need algorithms to learn non-linear features in data
- Examples: support vector machines, k-nearest neighbours, decision trees, ensemble methods ...

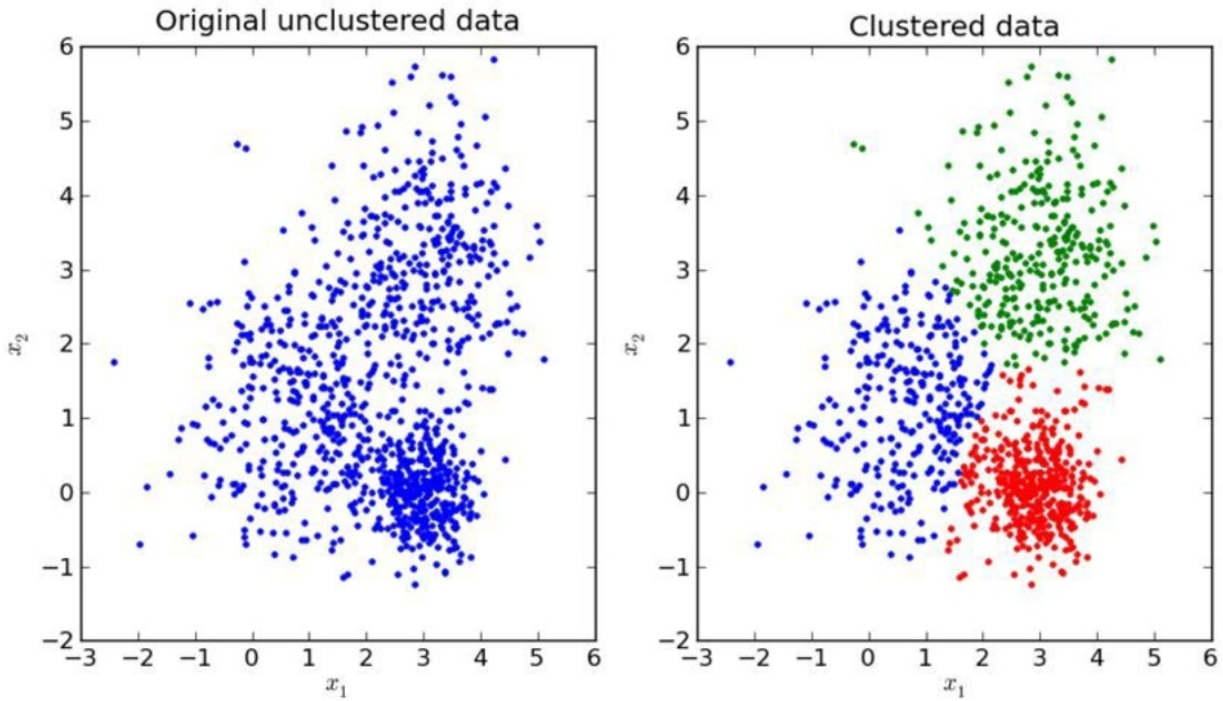


8

[8] https://scikit-learn.org/stable/auto_examples/svm/plot_svm_nonlinear.html

Unsupervised learning

- Datasets have no labels - no supervision
- Extract structures in datasets
- Unsupervised approaches - clustering, dimensionality reduction, ...

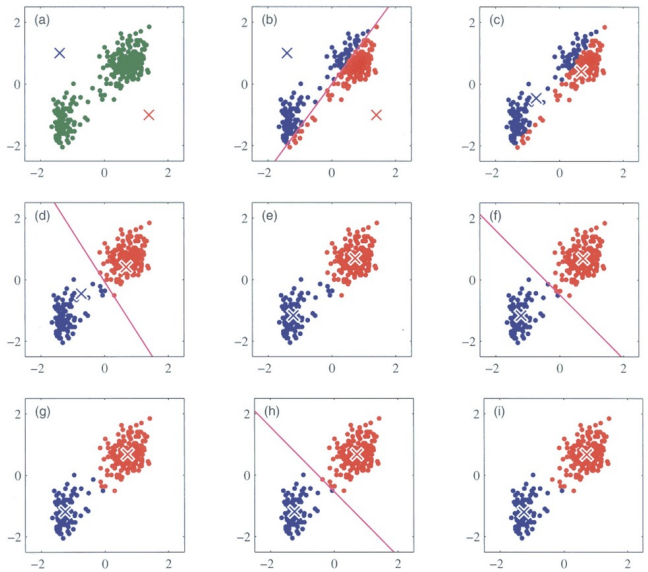


9

[9] https://deepai.org/machine-learning-glossary-and-terms/unsupervised-learning

Clustering

- Group data points based on similarity
- Similarity is determined by a notion of closeness
- Iterative process
- Types of clustering: k-means, hierarchical clustering, density-based spatial clustering of applications with noise (DBSCAN)

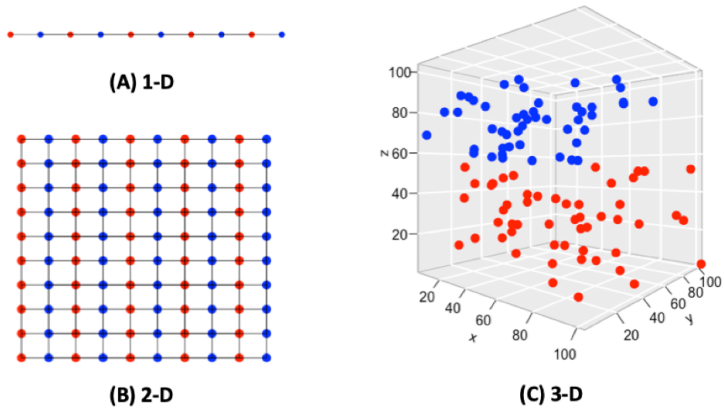


10

[10] http://dendroid.sk/2011/05/09/k-means-clustering/

Dimensionality reduction

- Number of dimensions >> number of samples
- High dimensional dataset - data points become farther from one another
- Curse of dimensionality - hard to generalise on all combinations of a large number of dimensions
- May lead to high-variance or overfitting
- Remedy - remove noisy or insignificant dimensions
- Approaches: principal component analysis (PCA), autoencoders (Neural network)

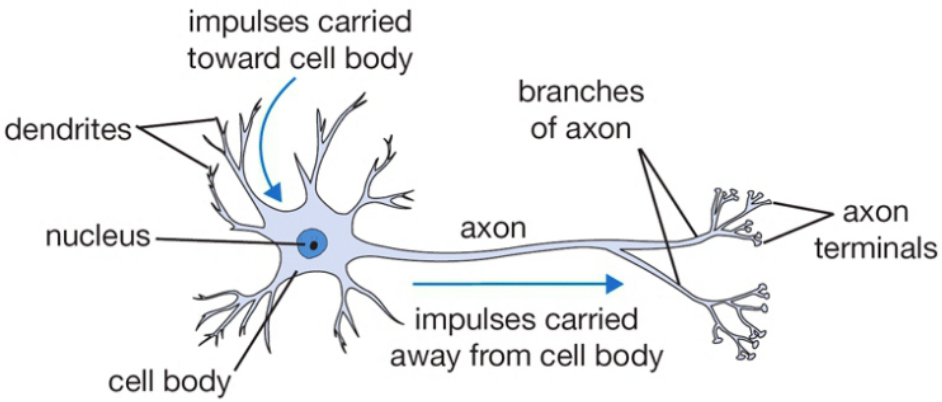


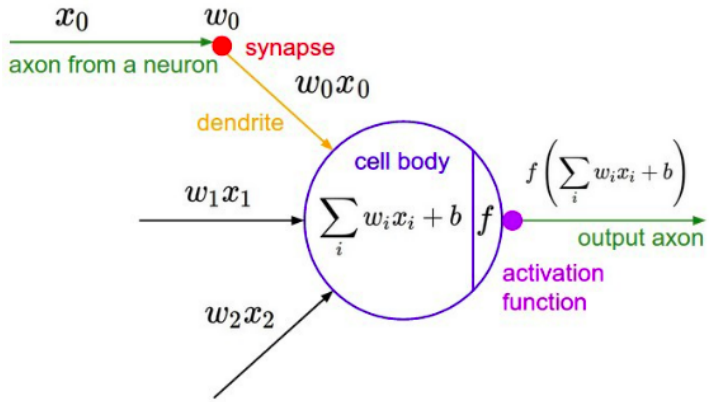
11

[11] https://cofactorgenomics.com/curse-of-dimensionality-wk-16/

Artificial neural networks

- Inspired by biological neurons
- Dendrites and axons carry signals
- In an artificial neural network, neural network edges carry data to and from neurons



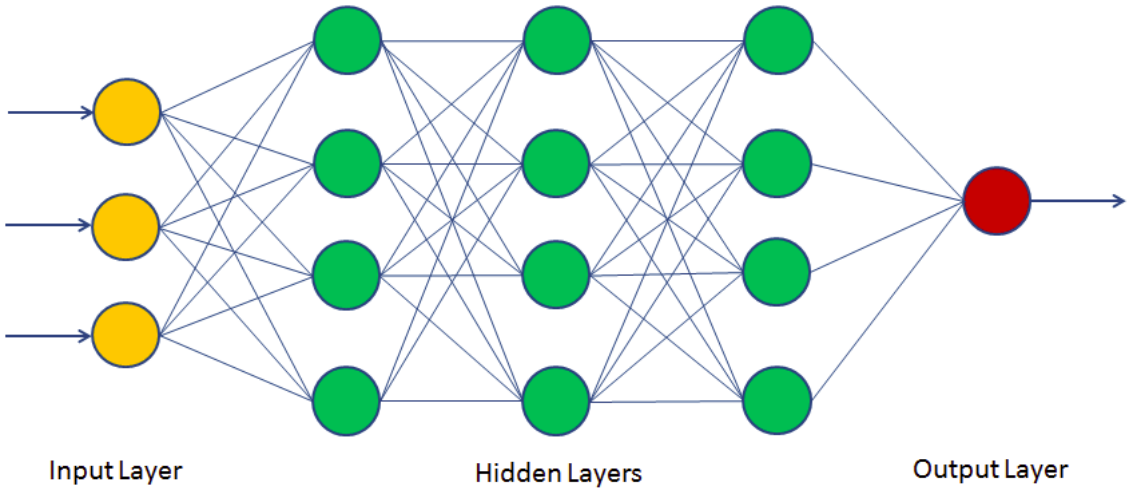


12

[12] https://towardsdatascience.com/a-gentle-introduction-to-neural-networks-series-part-1-2b90b87795bc

Artificial neural networks

- Neural network has an architecture
- Layers - input, hidden, output,
- Loss function - mean squared error, cross-entropy loss.
- Optimiser - adam, adadelta, rmsprop ...
- Types of layers - recurrent, dropout, convolutional ...
- Types of activations - tanh, sigmoid, softmax ...

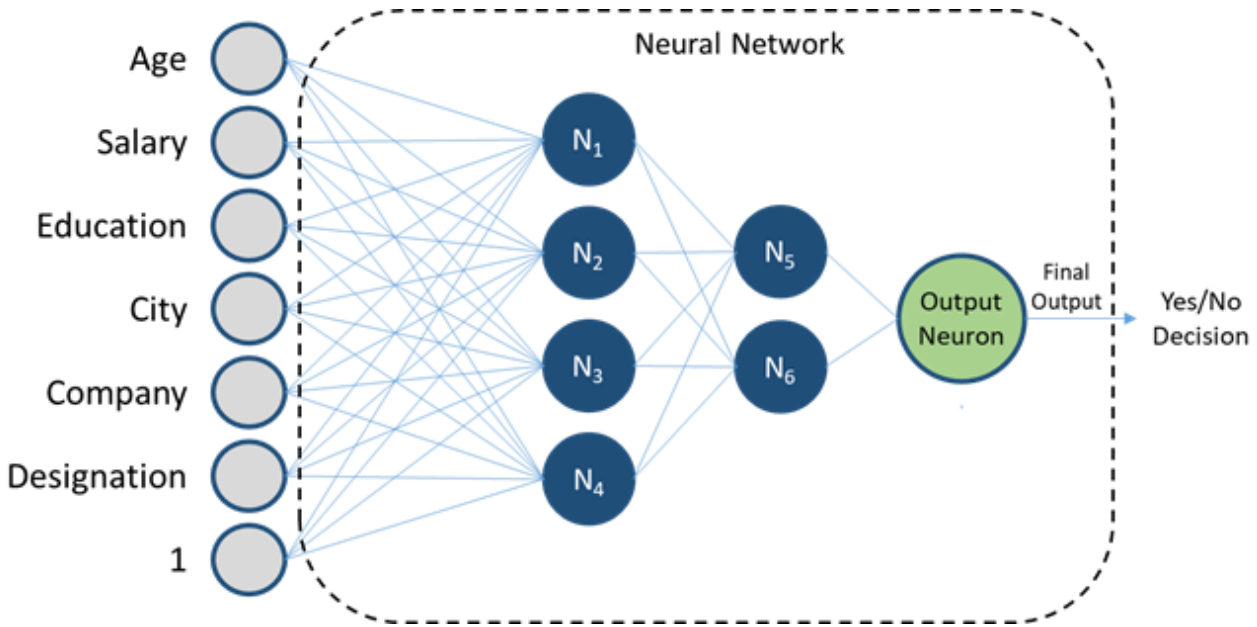


13

[13] https://www.datacamp.com/community/tutorials/neural-network-models-r

Artificial neural networks

- Input layer - receive data
- Number of neurons = number of features
- Hidden layer - number of neurons or layers not fixed, depends on the problem being solved. Responsible for learning complex patterns
- Output layer - compute output as a class or real number



14

[14] https://www.datacamp.com/community/tutorials/neural-network-models-r

General recommendations for using ML algorithms

- Preprocess datasets - outliers, incorrect labels, standardise features by scaling, encoding, imputing missing values
- Split datasets - train, test, validation and K-fold cross-validation
- Use right algorithm - start with simple and then move to complex
- Fix data imbalance
- Tune hyperparameters
- Look for overfitting
- Evaluate accuracy for each class (for classification)

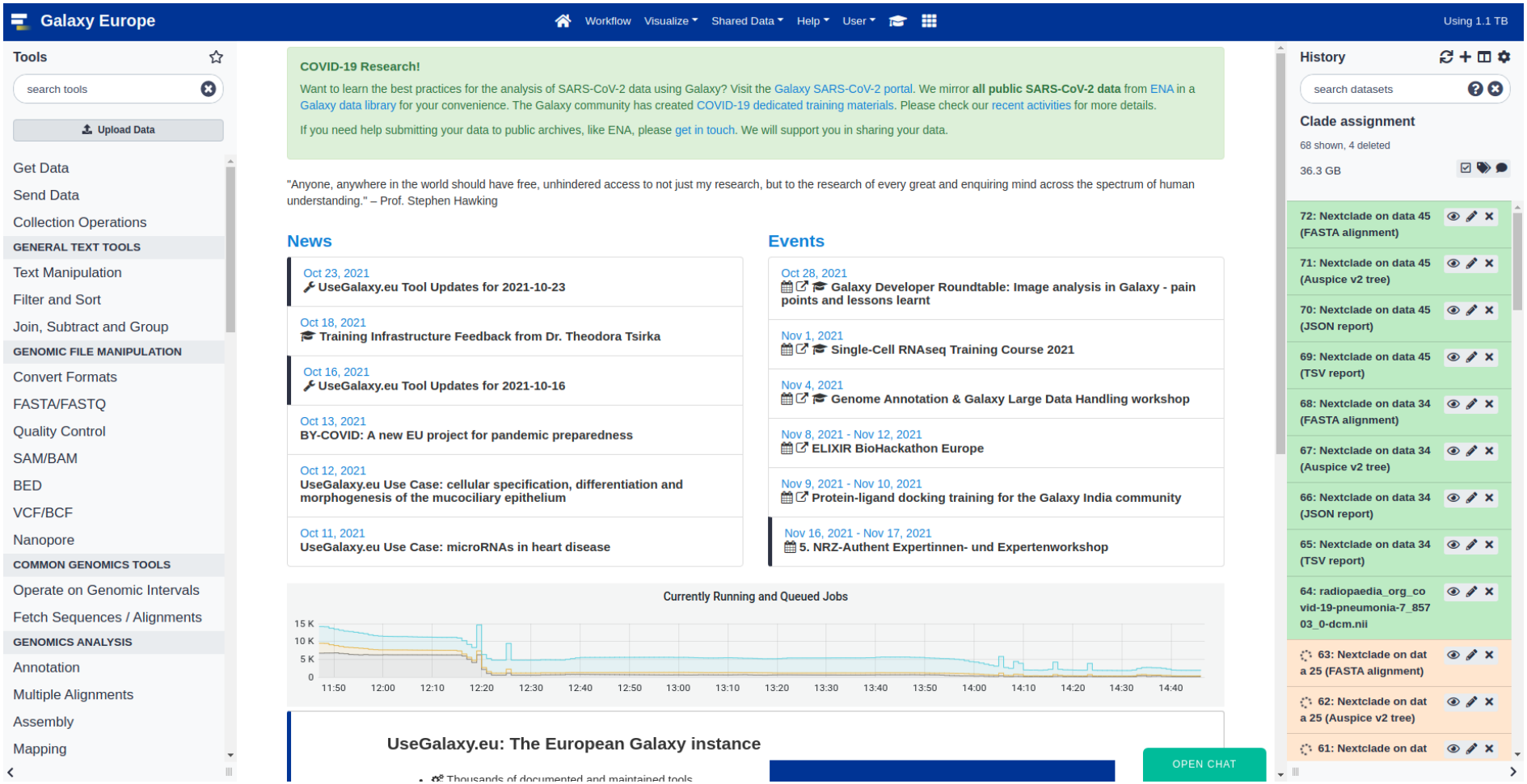
Machine learning in Galaxy

Galaxy Europe

- Online platform for numerous (>2000) scientific tools running on large compute resources including GPUs as well as large storage
- Accelerates scientific, especially bioinformatics research
- Public infrastructure
- Open-source community, contributors across the globe
- Over 200 tutorials (hands-on materials) showing usage of tools in different scientific analyses [15]

[15] https://training.galaxyproject.org/

Galaxy Europe



16

[16] https://usegalaxy.eu/

ML in Galaxy

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Galaxy-ML: An accessible, reproducible, and scalable machine learning toolkit for biomedicine

Qiang Gu[1,2], **Anup Kumar**[3], **Simon Bray**[3], **Allison Creason**[1,2], **Alireza Khanteymoori**[3], **Vahid Jalili**[1,2], **Björn Grüning**[3], **Jeremy Goecks**[1,2]*

1 Department of Biomedical Engineering, Oregon Health & Science University, Portland, Oregon, United States of America, **2** The Knight Cancer Institute, Oregon Health & Science University, Portland, Oregon, United States of America, **3** Bioinformatics Group, Department of Computer Science, University of Freiburg, Freiburg, Germany

* goecksj@ohsu.edu

OPEN ACCESS

Citation: Gu Q, Kumar A, Bray S, Creason A, Khanteymoori A, Jalili V, et al. (2021) Galaxy-ML: An accessible, reproducible, and scalable machine learning toolkit for biomedicine. PLoS Comput Biol 17(6): e1009014. https://doi.org/10.1371/journal.pcbi.1009014

Abstract

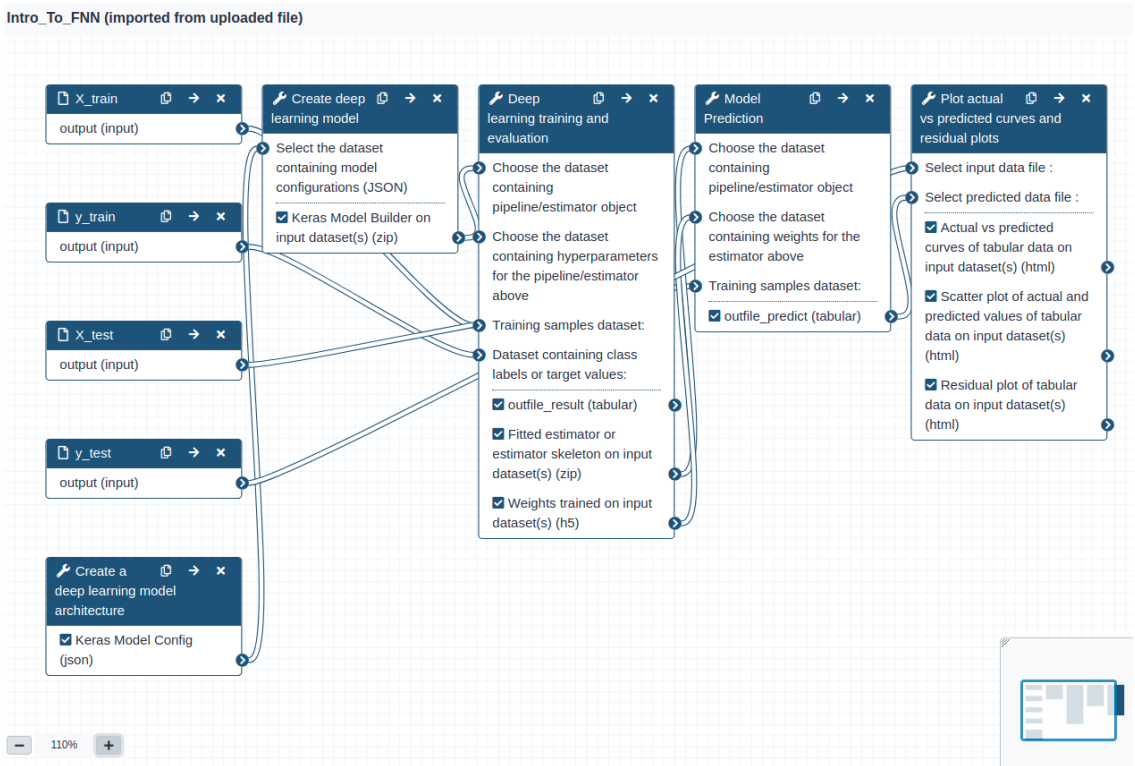
Supervised machine learning is an essential but difficult to use approach in biomedical data analysis. The Galaxy-ML toolkit (https://galaxyproject.org/community/machine-learning/) makes supervised machine learning more accessible to biomedical scientists by enabling them to perform end-to-end reproducible machine learning analyses at large scale using only a web browser. Galaxy-ML extends Galaxy (https://galaxyproject.org), a biomedical computational workbench used by tens of thousands of scientists across the world, with a suite of tools for all aspects of supervised machine learning.

17

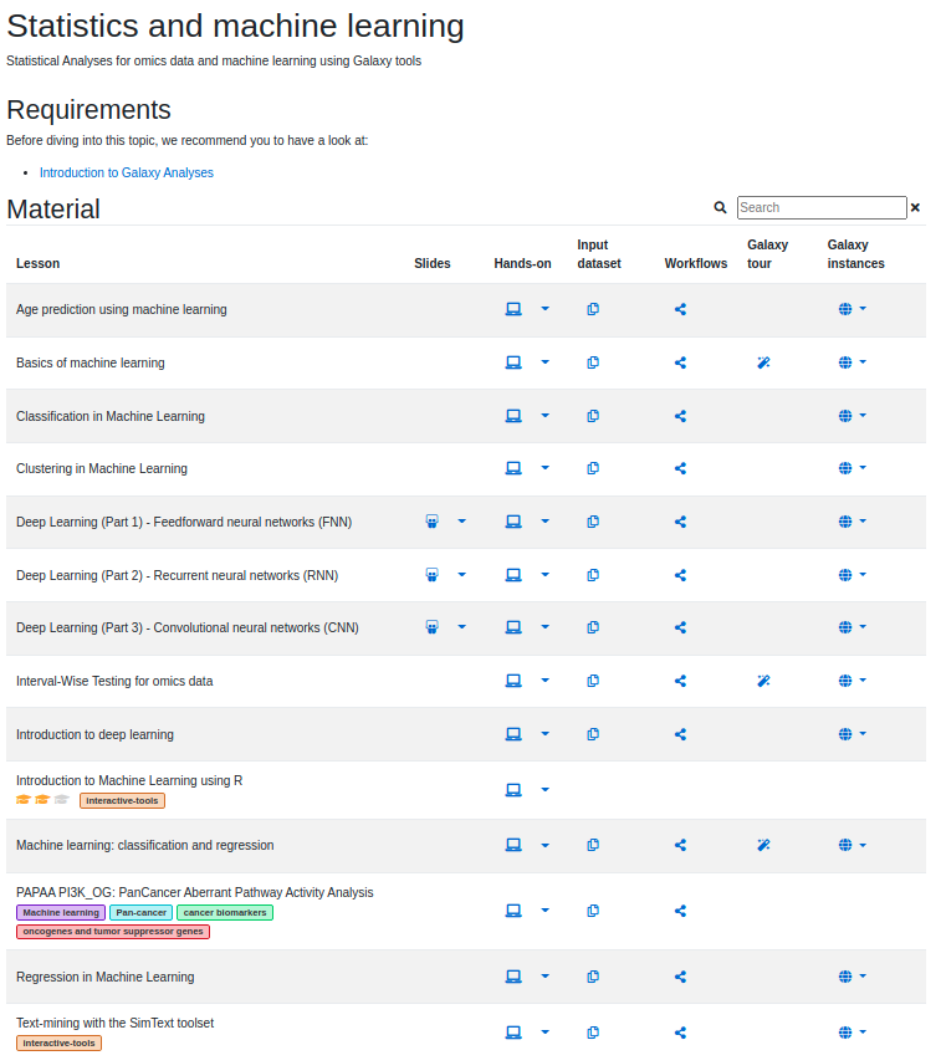
[17] https://doi.org/10.1371/journal

ML in Galaxy

- 20 - 30 ML tools powered by scikit-learn and tensorflow
- ML tools - classifiers, regressors, data preprocessors, visualizations, hyperparameter tuners, pipeline builders
- Workflow of tools
- Long running ML training on Galaxy infrastructure (using multiple CPUs, GPUs)



ML tutorials in Galaxy



Statistics and machine learning

Statistical Analyses for omics data and machine learning using Galaxy tools

Requirements

Before diving into this topic, we recommend you to have a look at:

- Introduction to Galaxy Analyses

Material

Search

Lesson	Slides	Hands-on	Input dataset	Workflows	Galaxy tour	Galaxy instances
Age prediction using machine learning		✓	✓	✓		✓
Basics of machine learning		✓	✓	✓	✓	✓
Classification in Machine Learning		✓	✓	✓		✓
Clustering in Machine Learning		✓	✓	✓		✓
Deep Learning (Part 1) - Feedforward neural networks (FNN)	✓	✓	✓	✓		✓
Deep Learning (Part 2) - Recurrent neural networks (RNN)	✓	✓	✓	✓		✓
Deep Learning (Part 3) - Convolutional neural networks (CNN)	✓	✓	✓	✓		✓
Interval-Wise Testing for omics data		✓	✓	✓	✓	✓
Introduction to deep learning		✓	✓	✓		✓
Introduction to Machine Learning using R interactive-tools		✓				
Machine learning: classification and regression		✓	✓	✓	✓	✓
PAPAA PI3K_OG: PanCancer Aberrant Pathway Activity Analysis Machine learning Pan-cancer cancer biomarkers oncogenes and tumor suppressor genes		✓	✓	✓		
Regression in Machine Learning		✓	✓	✓		✓
Text-mining with the SimText toolset interactive-tools		✓	✓	✓		✓

19

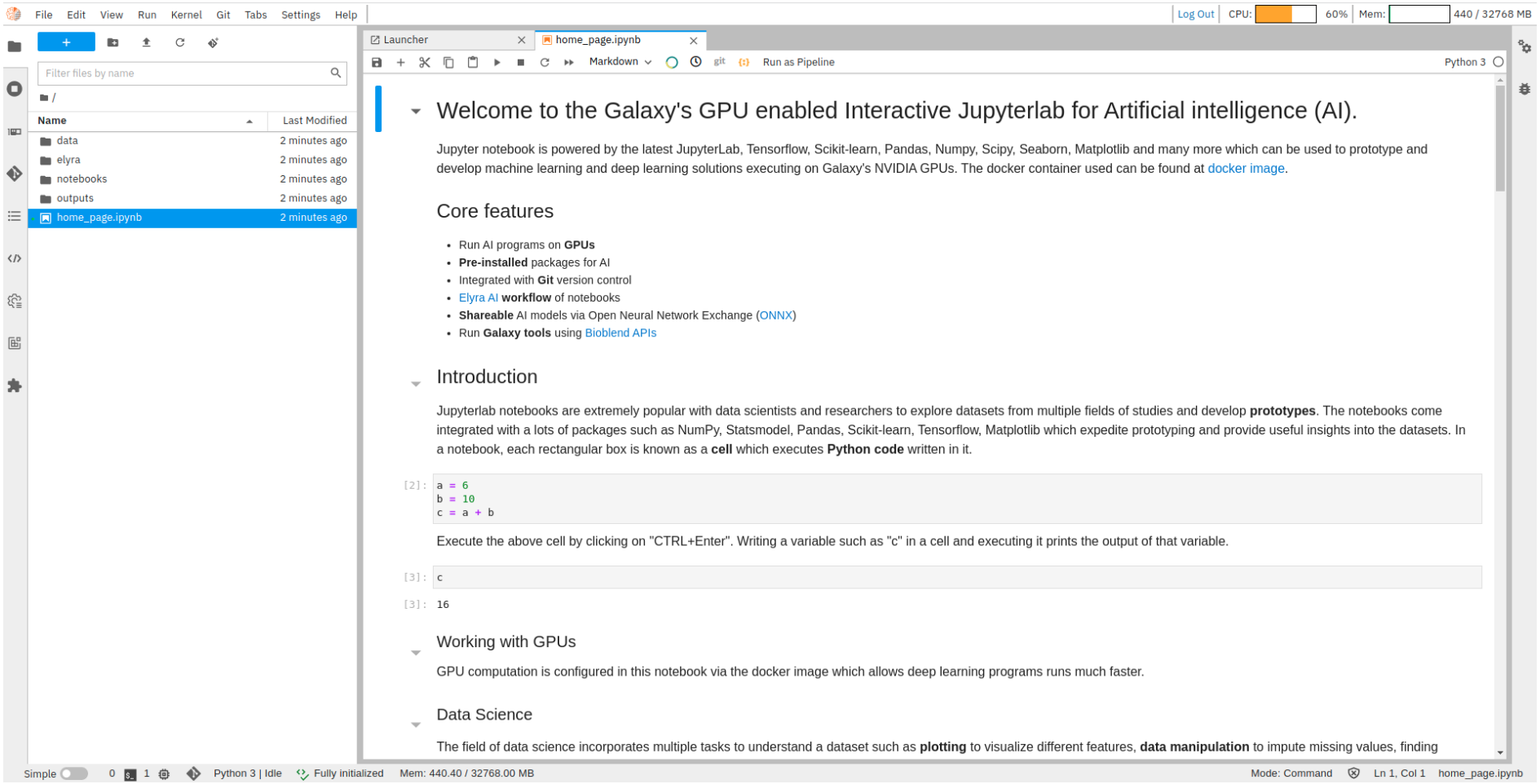
[19] https://training.galaxyproject.org/training-material/topics/statistics/

Our projects with machine learning

Galaxy Jupyterlab editor for ML

- Jupyter notebook - popular editor
- Scientific computing, data science, machine learning, learn to code ...
- Simple and fast way to create prototypes
- No need for any package installation
- Easy to share any analysis
- Runs on web

Galaxy Jupyterlab [20]



[20] https://live.usegalaxy.eu/?tool_id=interactive_tool_ml_jupyter_notebook

Features of Galaxy Jupyterlab

- Base container - jupyter/tensorflow-notebook:latest [21]
- CUDA and cuDNN packages for nvidia GPUs, tensorflow for GPU, pre-installed ML and DL packages
- Create, share and reuse ML/DL models - ONNX [22]
- Git integration
- Workflow of notebooks - Elyra AI [23]
- Connect to Galaxy histories, datasets using bioblend [24]
- Miscellaneous - dashboards for CPU, GPU, memory utilization, collapse/expand sections, notebook as voila ...
- Docker image [25]

[21] https://hub.docker.com/r/jupyter/tensorflow-notebook/

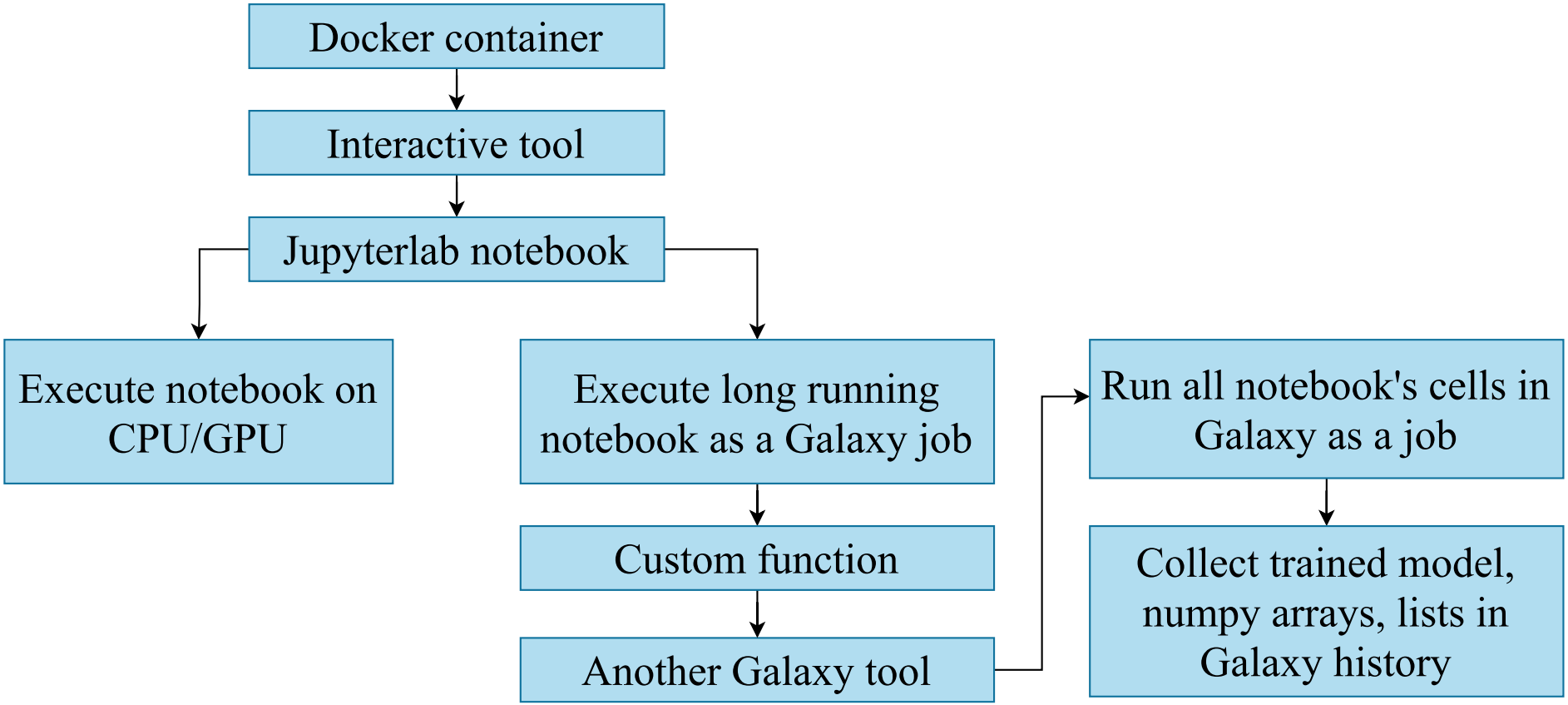
[22] https://onnx.ai/

[23] https://github.com/elyra-ai/elyra

[24] https://bioblend.readthedocs.io/en/latest/

[25] https://github.com/anuprulez/ml-jupyter-notebook

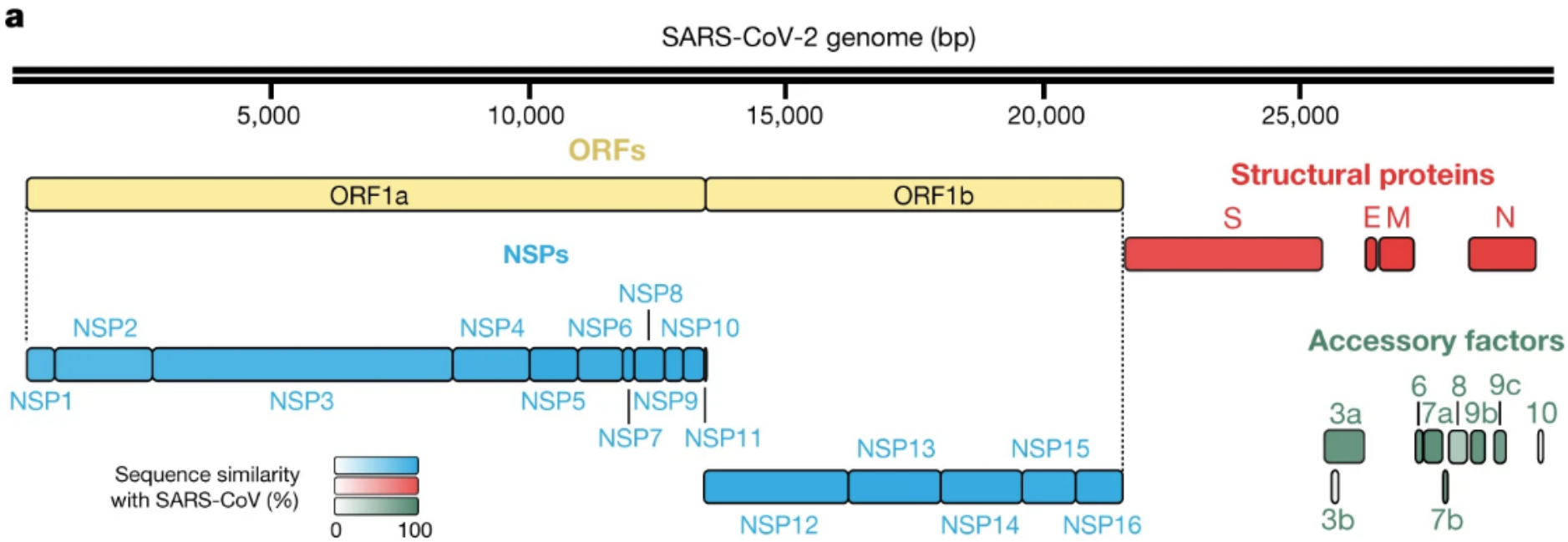
Architecture of Galaxy Jupyterlab



Prediction of protein evolution (amino acid substitutions) in SARS-COV2 sequences

- Spike protein and amino acid (AA) mutations
- Nextclade clades
- Sequence to sequence learning
- Generative adversarial networks (GANs)
- Comparison of true and generated SARS-CoV-2 AA sequences
- Substitutions from generated AA sequences (future substitutions)?

Spike protein (S)

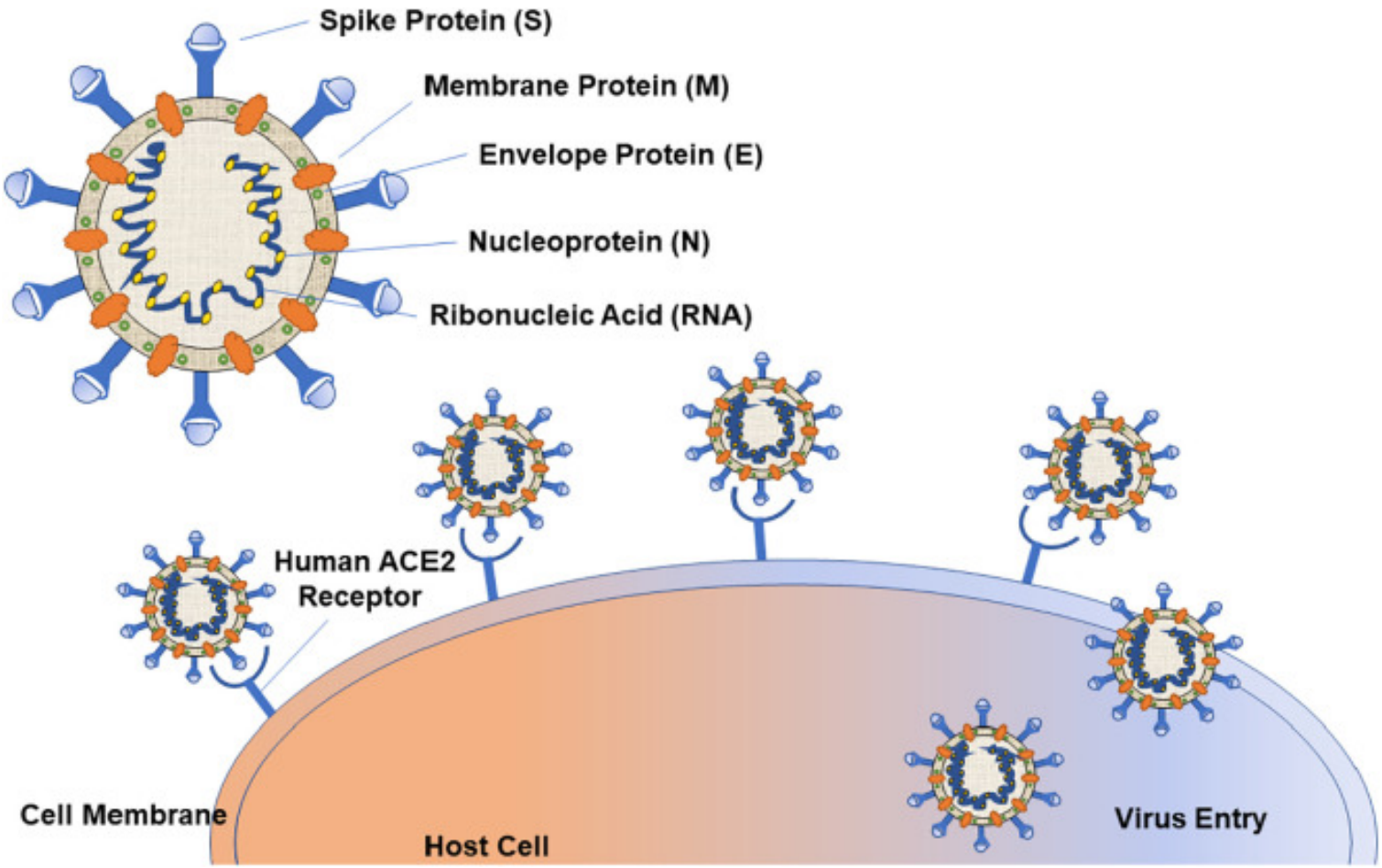


26

- Non-structural and structural proteins

[26] https://www.nature.com/articles/s41586-020-2286-9

Spike protein (S)



27

[27] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7293463/

Spike protein (S)

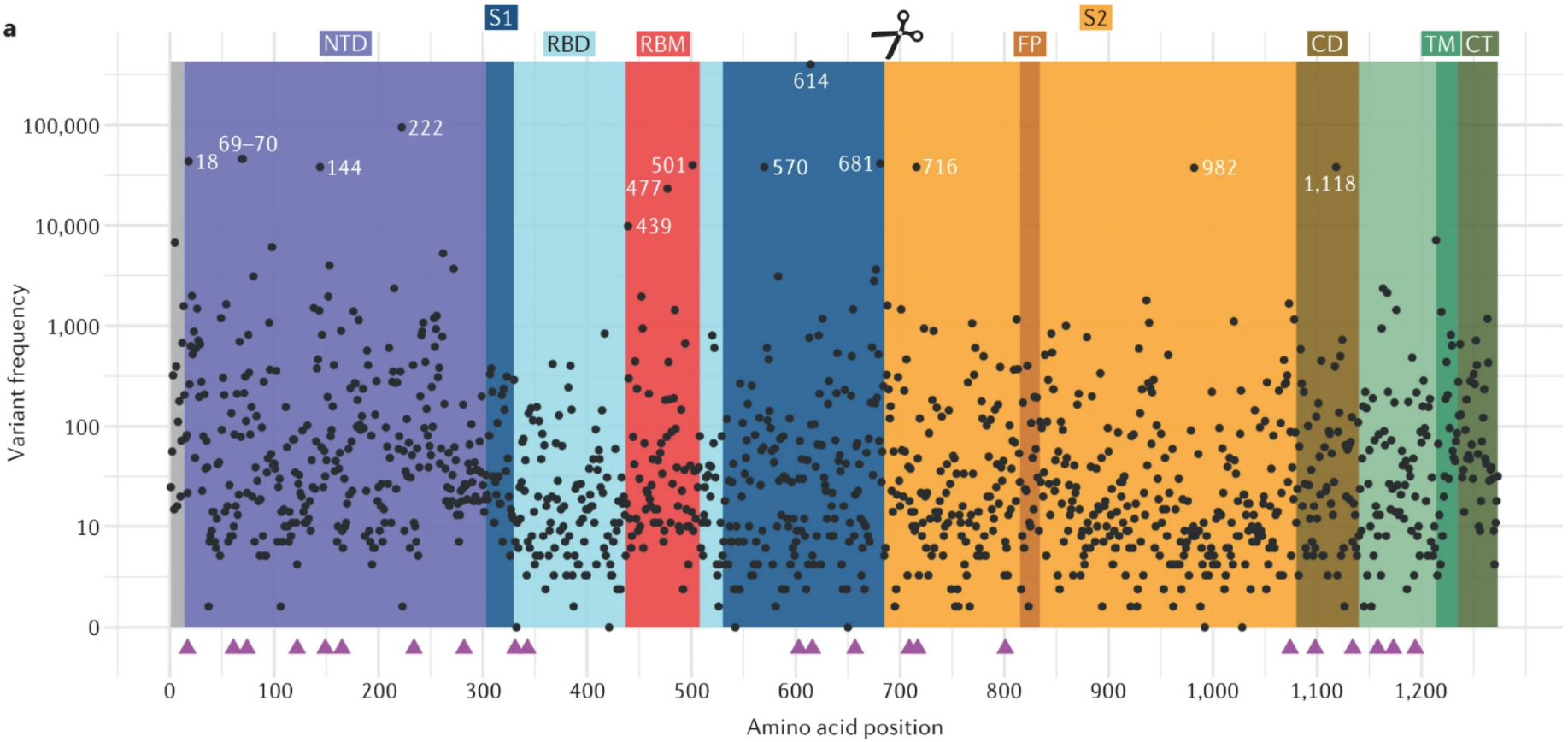
- Binds to the host cell
- Mutations may impact infectivity, transmissibility
- D614G: enhances viral replication [28]
- N439K: enhances the binding affinity for the ACE2 receptor and reduces the neutralizing activity of antibodies [29]
- Y453F: increased ACE2-binding affinity [30]
- ...

[28] https://www.nature.com/articles/s41586-020-2895-3, https://covariants.org/variants/20B.S.732A

[29] https://www.nature.com/articles/s41579-021-00573-0

[30] https://www.nature.com/articles/s41579-021-00573-0

Frequency of spike mutations (substitutions and deletions)

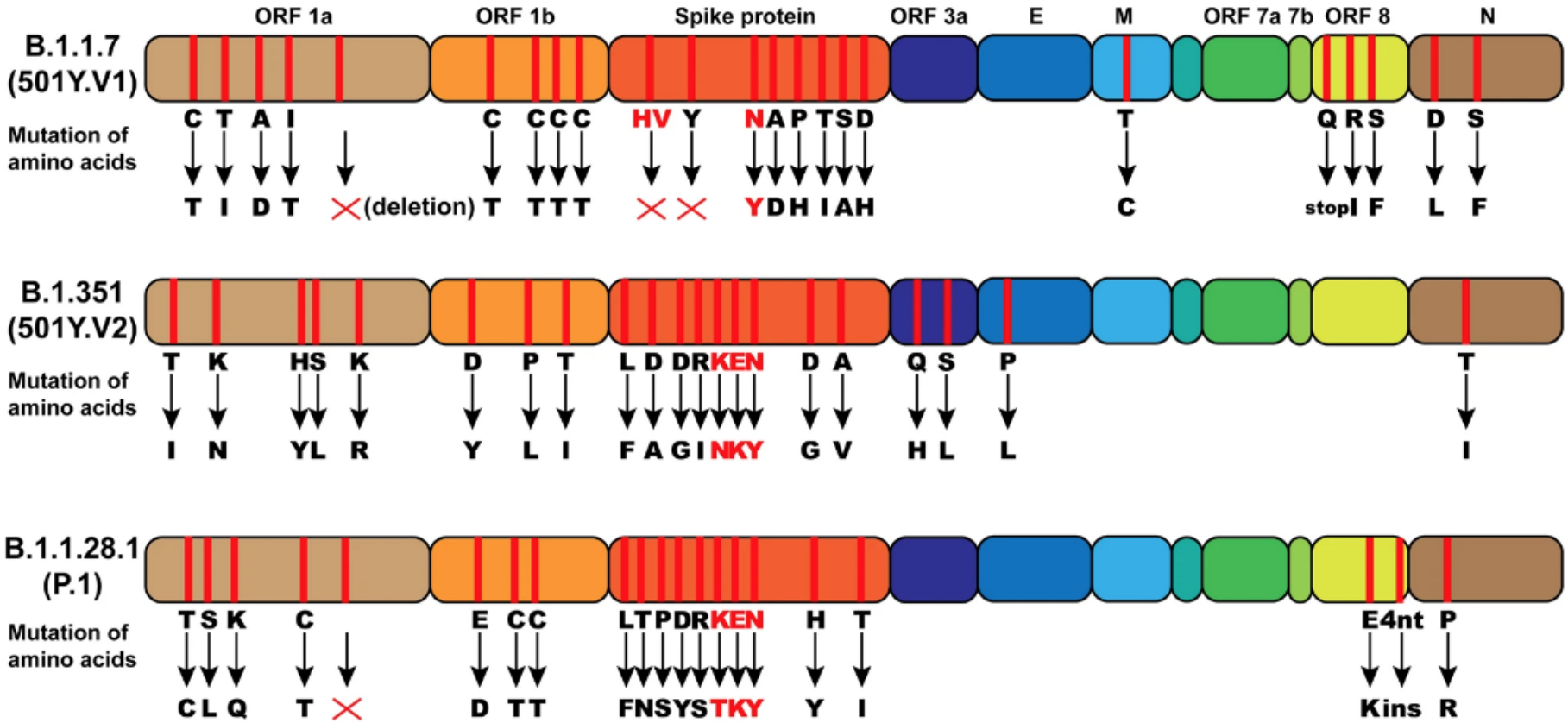


31

- 426,623 genomes, 5106 substitutions

[31] https://www.nature.com/articles/s41579-021-00573-0

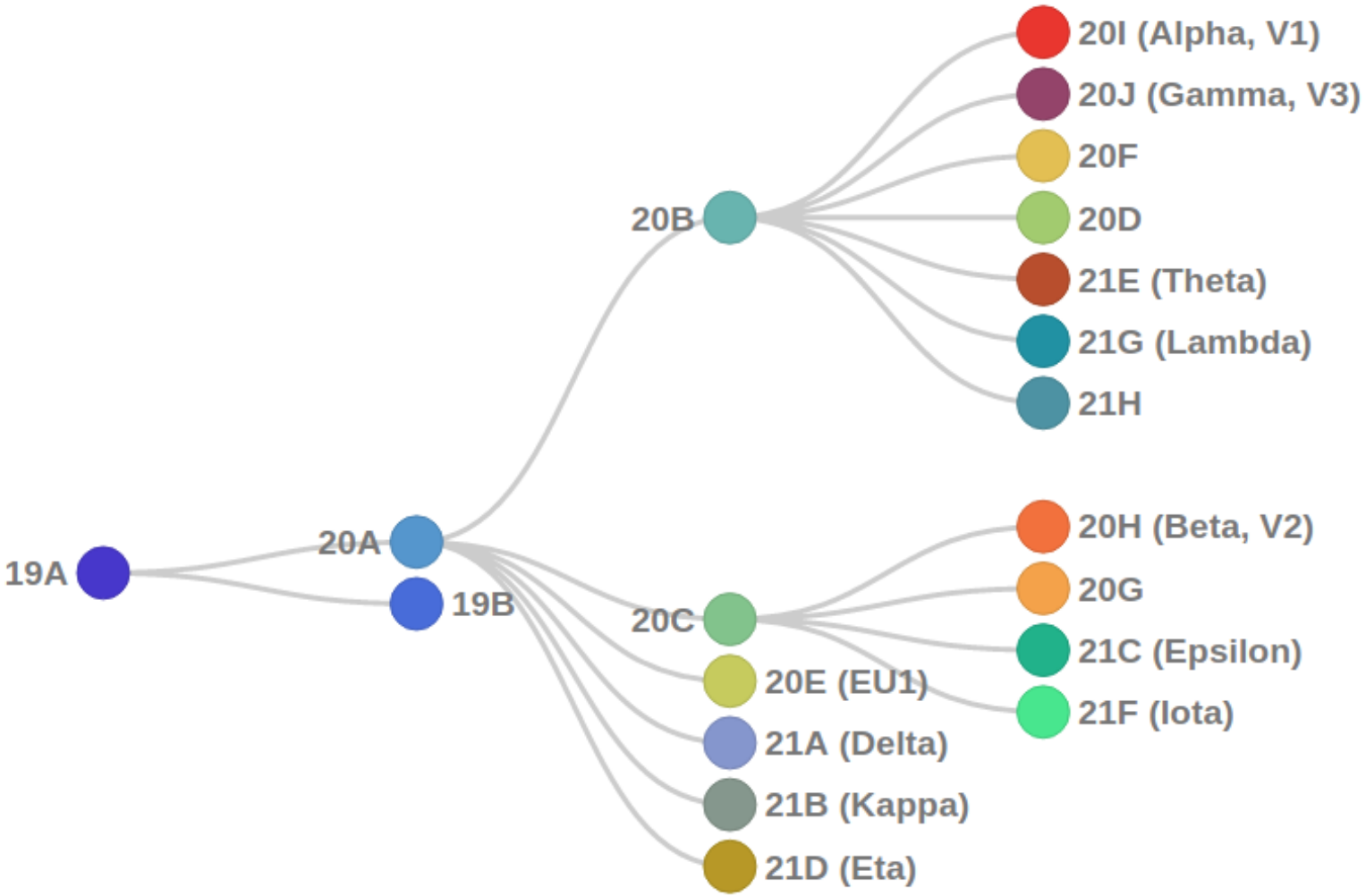
Spike mutations in lineages



32

[32] https://www.nature.com/articles/s41392-021-00644-x

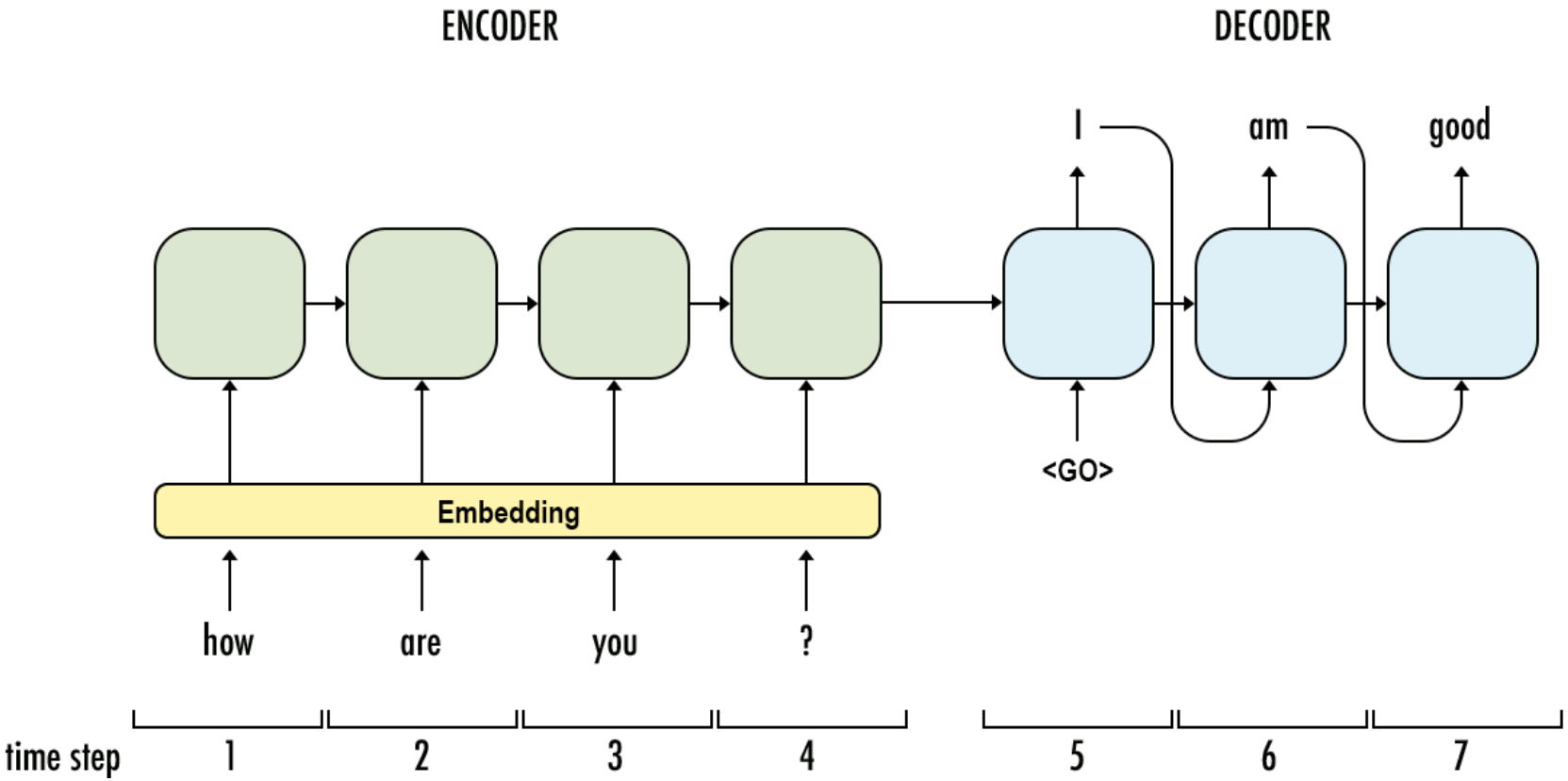
Nextclade clades



33

[33] https://clades.nextstrain.org/

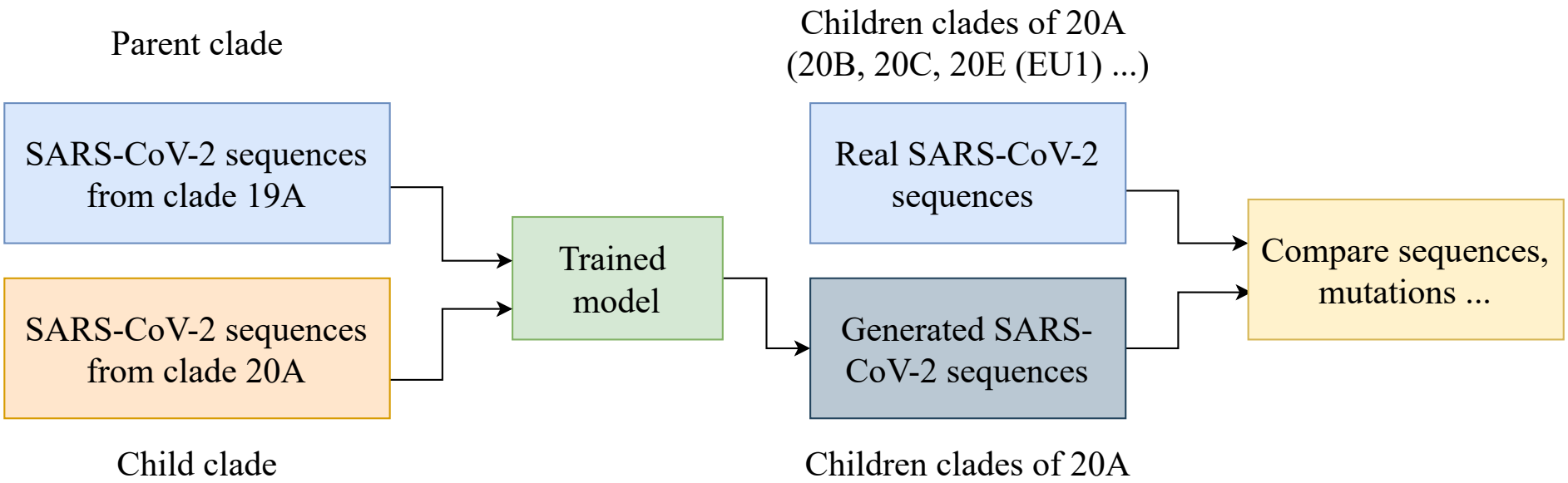
Sequence to sequence learning



34

[34] https://towardsdatascience.com/sequence-to-sequence-model-introduction-and-concepts-44d9b41cd42d

Sequence to sequence learning with SARS-CoV-2 sequences



Generative Adversarial Networks (GANs)

- Generator - generates data (sequences)
- Generator network - sequence to sequence encoder-decoder network
- Discriminator - discriminates between real and generated data (sequences)
- Discriminator network - sequential network to predict either real (true) or generated (false)
- Generator and Discriminator - make each other better over training iteration
- Applications - improve astronomical images [35], reconstruct 3D model of object from images [36], age face photographs [37], language translation[38], ...

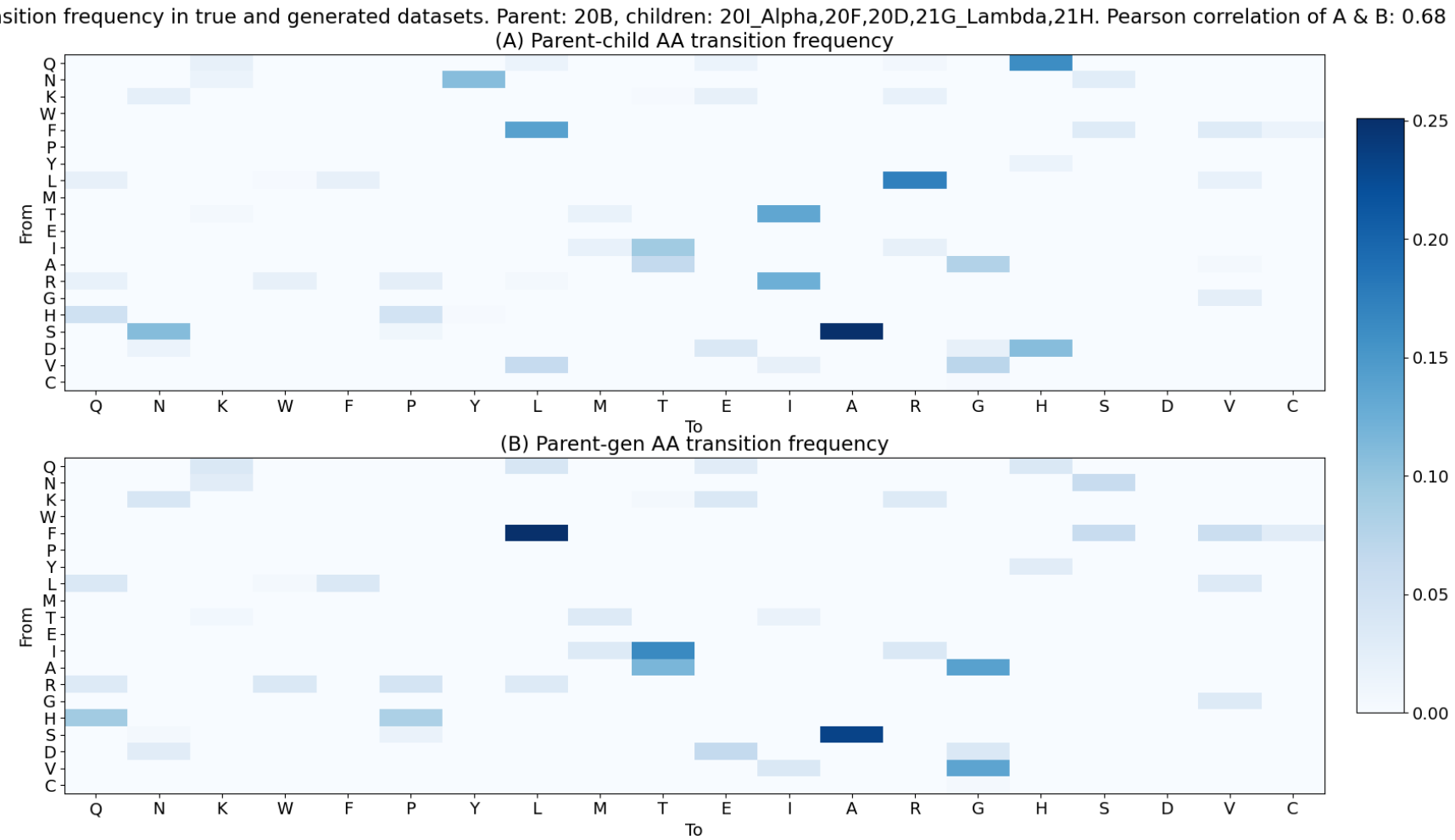
[35] https://arxiv.org/pdf/1702.00403.pdf

[36] http://3dgan.csail.mit.edu/

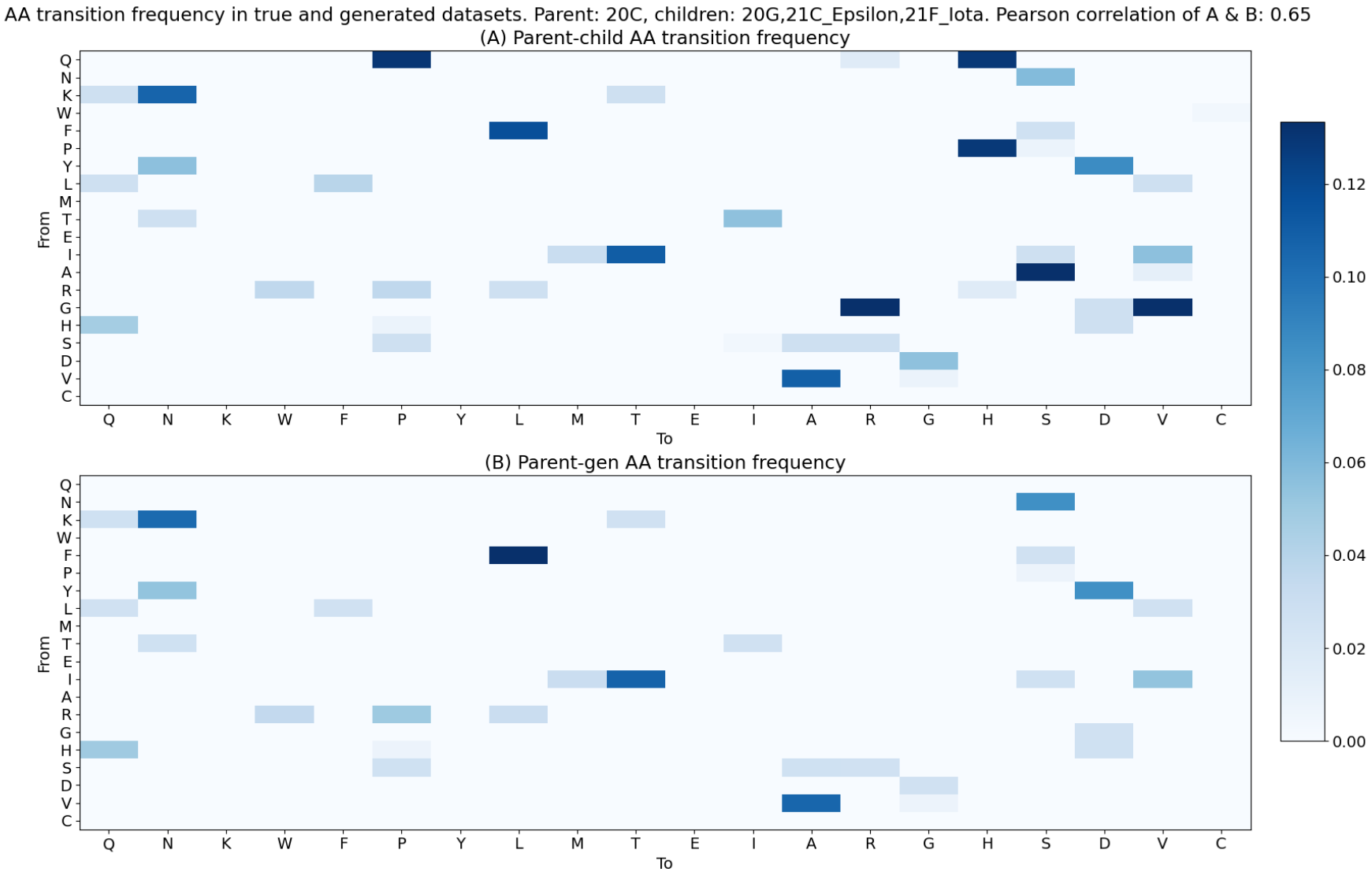
[37] https://arxiv.org/pdf/1702.01983.pdf

[38] https://arxiv.org/pdf/1704.06933.pdf

Prediction of protein evolution (for children clades of 20B)



Prediction of protein evolution (for children clades of 20C)



Thank you for your attention. Questions?