THE ELIXIR MACHINE LEARNING FOCUS GROUP: ACHIEVEMENTS AND ROAD AHEAD

Fotis Psomopoulos
Institute of Applied Biosciences, Centre for Research and Technology Hellas

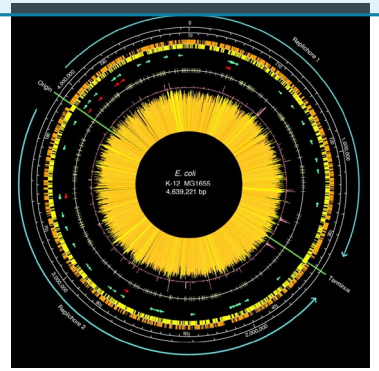
CC BY SA

INSTITUTE OF APPLIED BIOSCIENCES
ΙΝΣΤΙΤΟΥΤΟ ΕΦΑΡΜΟΣΜΕΝΩΝ ΒΙΟΕΠΙΣΤΗΜΩΝ
CENTRE for RECEARCH and TECHNOLOGY-HELLAS

CERTH
CENTRE FOR RESEARCH & TECHNOLOGY HELLAS

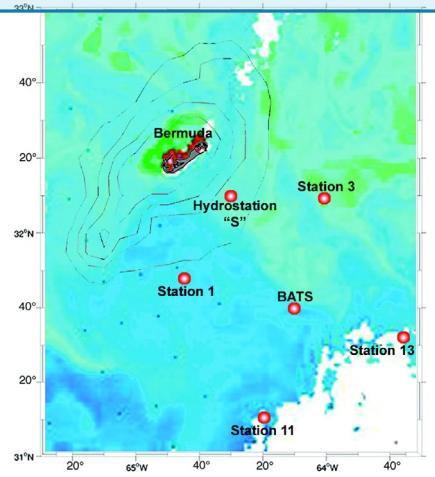
FROM GENOMICS TO META-GENOMICS

E. coli, Science, 1997

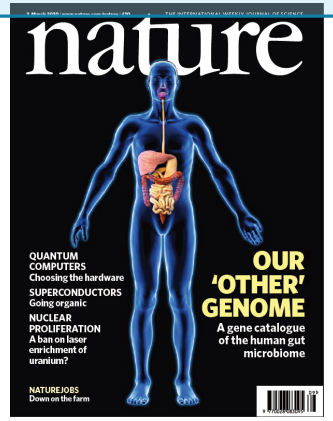


Human, Nature/Science, 2001

Saragasso sea, Science, 2004



Human gut, Nature, 2010

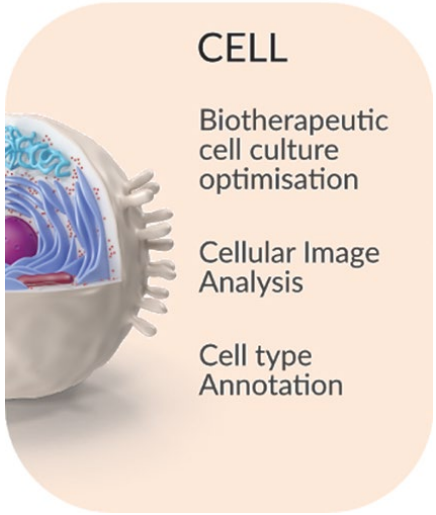


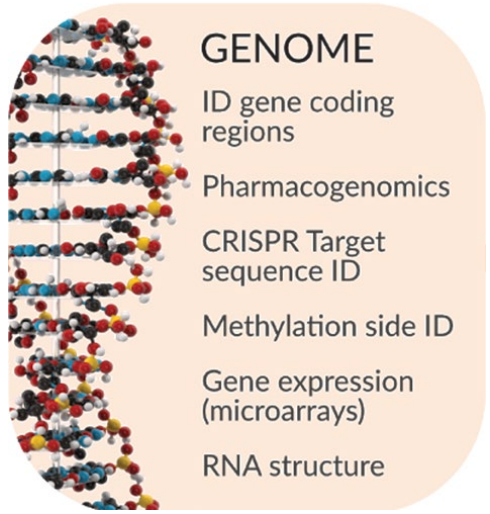
BASICALLY TRYING TO MATCH LEGO BLOCKS THAT FIT TOGETHER (GENOMIC JIGSAW).

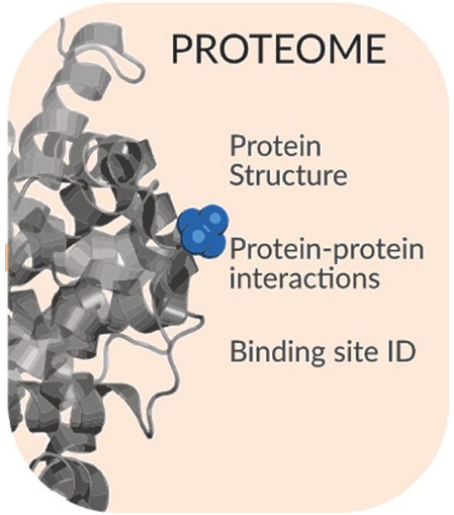
Easy right?

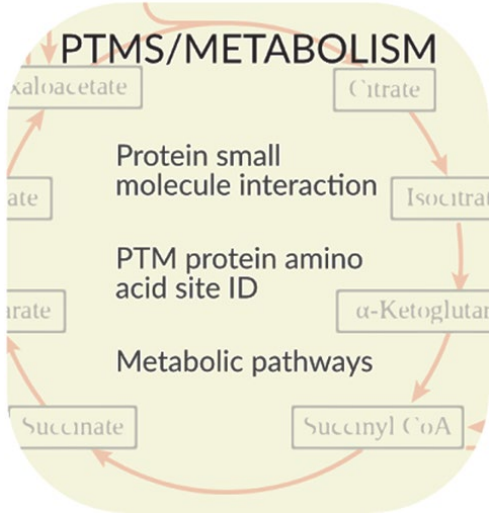
CC BY SA

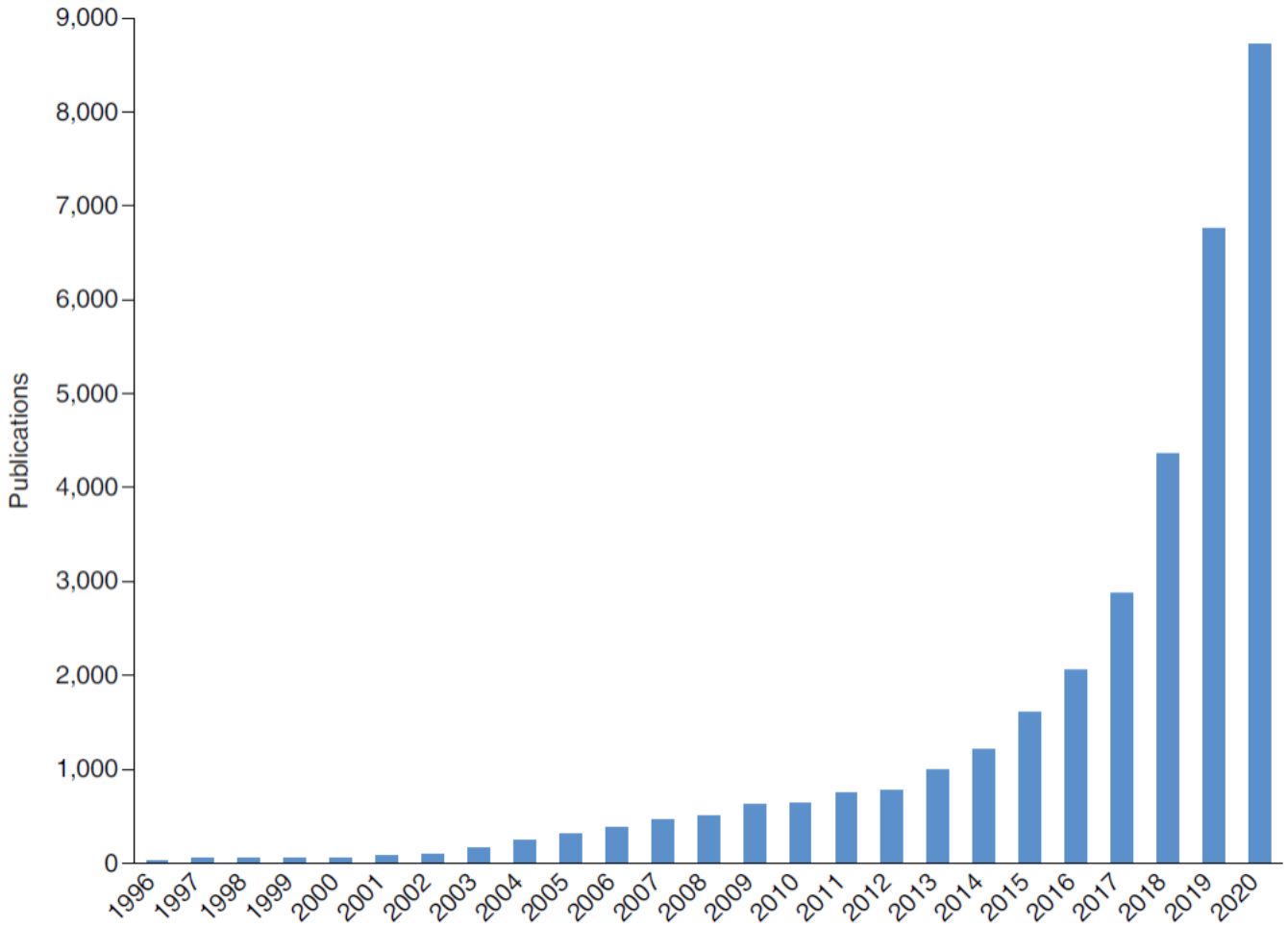
Machine learning is being used across many biological areas











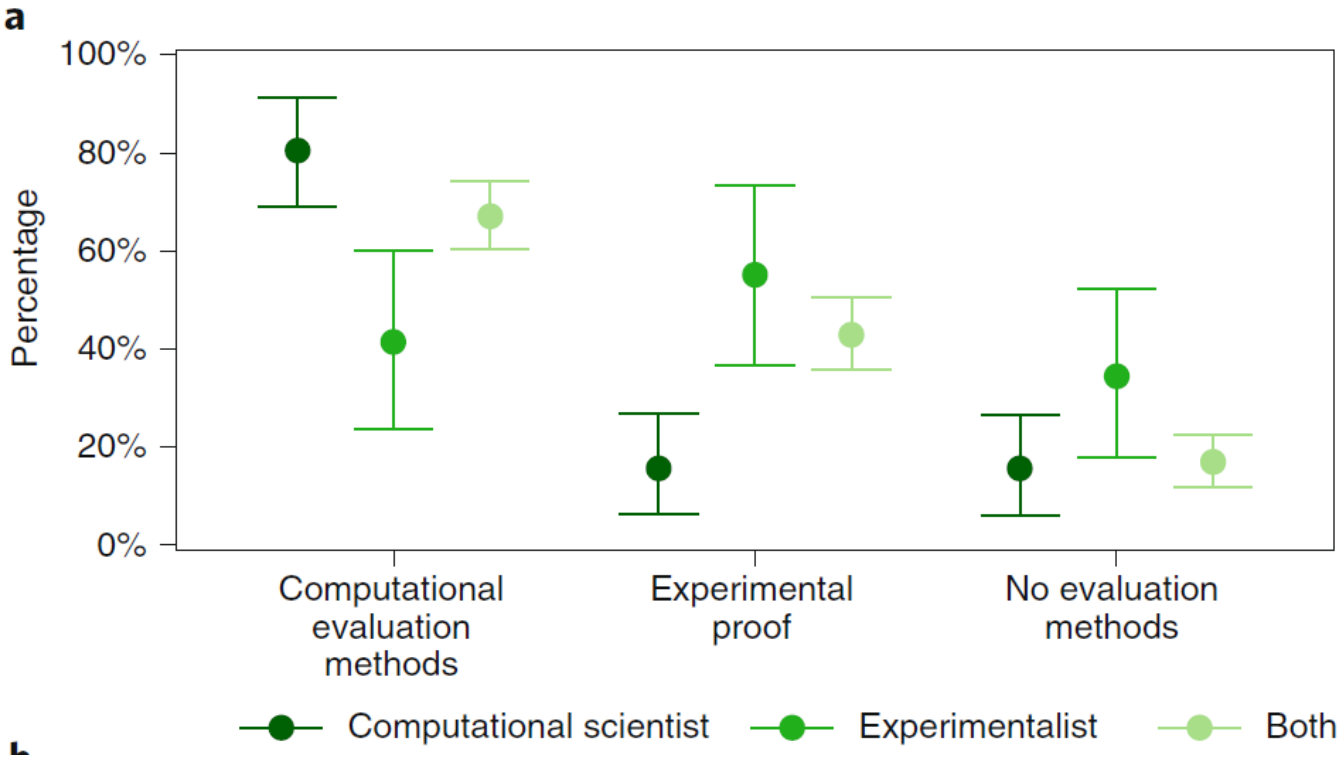
Corresponding **growth** of ML publications

The number of ML publications per year is based on Web of Science from 1996 onwards using the topic category for "machine learning" in combination with each of the following terms: "biolog", "medicine", "genom*", "prote*", "cell*", "post translational", "metabolic" and "clinical".*

XX %

What is the percent of these publications that did not include any evaluation?

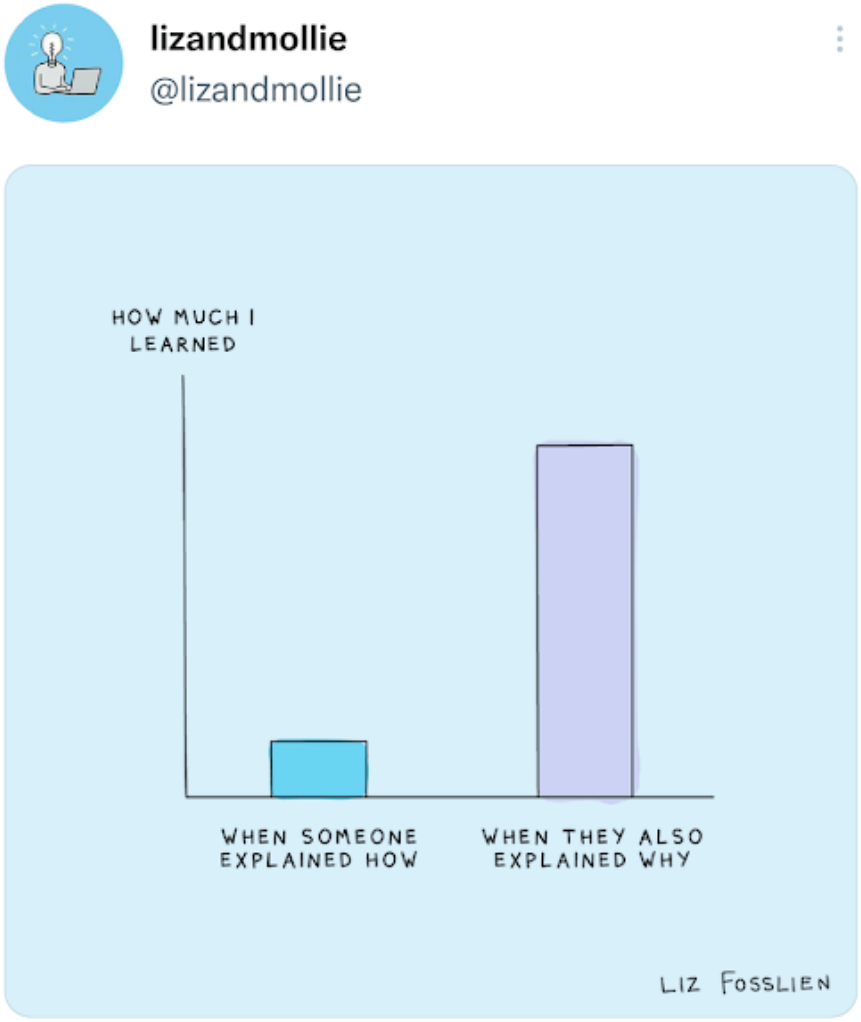
Go to www.menti.com and use the code 1259 0819



"~20% of publications did not apply any evaluation" [1]

[1] *Littmann, M. et al.* ***Validity of machine learning in biology and medicine increased through collaborations across fields of expertise****. Nat. Mach. Intell. 2, 18–24 (2020)*

methods and 43% provided experimental proof, suggesting that such collaborations facilitate experimental and computational validation. On the flip side, 19% of all articles did not provide any evaluation; this number rose as high as 34% without computational co-authors (Fig. 2a).



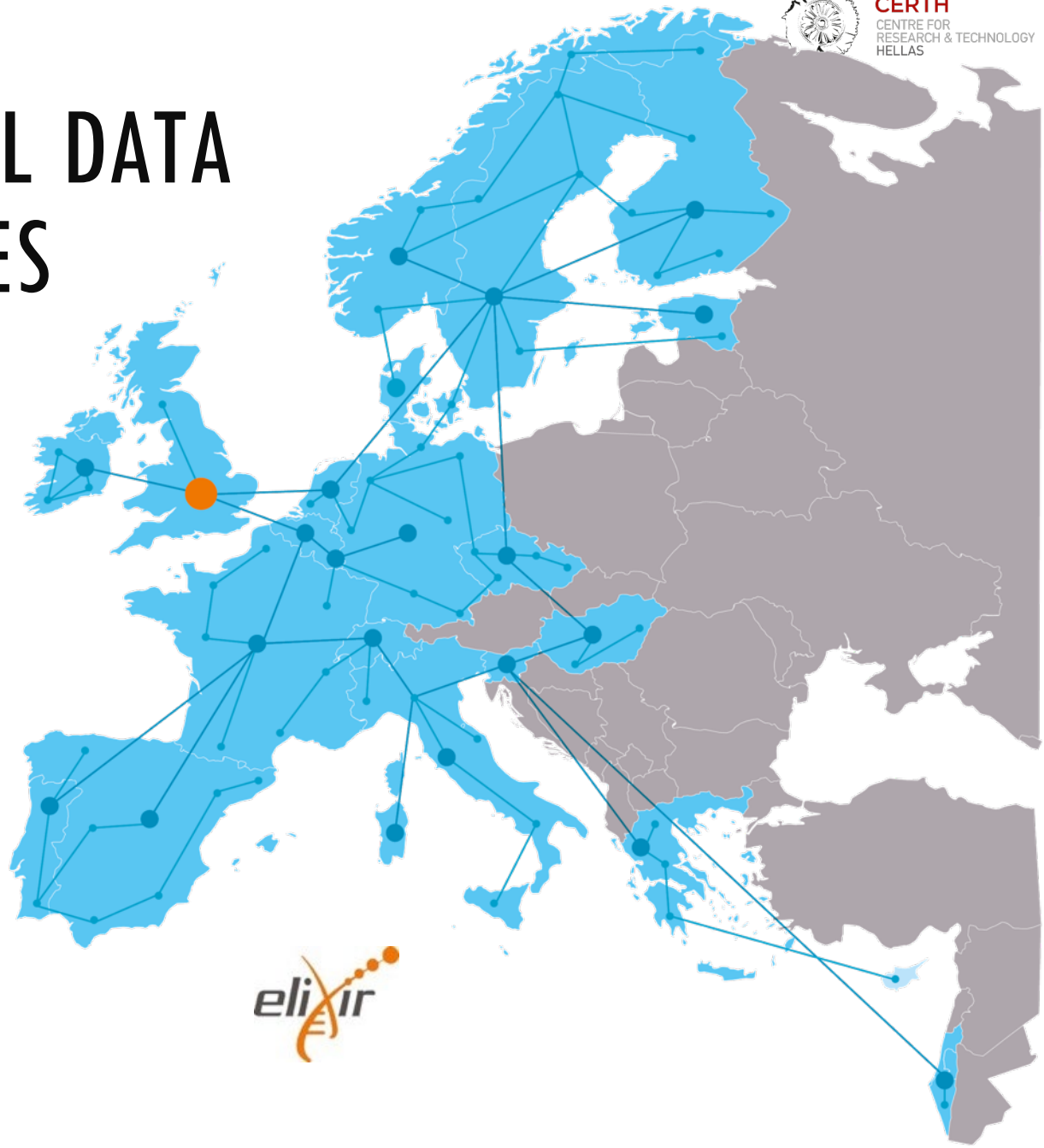
It’s not sufficient to present results;
you need all the details

Standards are the solution

CC BY SA

ELIXIR CONNECTS NATIONAL DATA NODES IN THE LIFE SCIENCES

- European infrastructure for life science data
- Align the services from national ELIXIR centres in a a federated ecosystem via common standards and services.
- Provide the people and technical capacity to enable FAIR data stewardship within every European life science project.



THE ML FOCUS GROUP IN A NUTSHELL

Jennifer Harrow
(ELIXIR Hub)

Fotis Psomopoulos
(ELIXIR Greece)

Silvio Tosatto
(ELIXIR Italy)

Goals

Standards for Machine Learning

Benchmarking of Machine Learning tools

Training for Machine Learning

Machine Learning **FAIR/reproducibility**

Integration across ELIXIR Communities

Wide interest across all Nodes

- consistent participation >40 members , representing >10 Nodes
- already one output (DOME Recommendations accepted in Nature Methods)
- maintaining a connection to Industry (Owkin, Pistoia alliance)

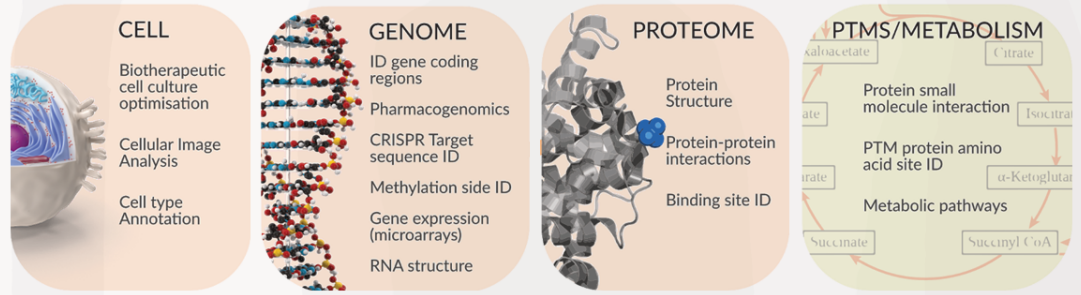
https://elixir-europe.org/focus-groups/machine-learning

A BRIEF HISTORY OF THE FOCUS GROUP

Jun 2019	First ad hoc discussions with various parties within ELIXIR interested in Machine Learning
Sep 2019	➢ Initial idea to set up the aims around the focus group ➢ David Jones comment on "Setting the standards for machine learning in biology" (*Nature Reviews Molecular Cell Biology*)
Oct 2019	Kick-off Meeting of the ML Focus Group (virtually, attracting ~30 participants)
Mar 2020	(5 months after) Submitted the DOME paper to Nature Methods (preprint on arXiv)
Apr 2020	Presented at the ELIXIR Bioinformatics Industry Forum 2020
Jun 2020	Presented at CLAIRE Task Force webinar on "AI & COVID-19: Results and next steps"
Jul 2020	Panel members at the Pistoia Alliance webinar for "Minimal Information About an AI Model"
Sep 2021	1st meeting of the Steering Committee, including people from academia, industry and publishers
Oct 2021	➢ DOME Recommendations in the Nature Methods October issue (editorial on keeping checks on machine learning, highlighting also DOME) ➢ Invited to present at the NIH FAIR Data WG Meeting

DOME RECOMMENDATIONS

Publication on DOME recommendations for life sciences published in nature methods



Set of recommendations for supervised learning with aim to improve standards

DOME: recommendations for supervised machine learning validation in biology

Ian Walsh, Dmytro Fishman, Dario Garcia-Gasulla, Tiina Titma, Gianluca Pollastri, ELIXIR Machine Learning Focus Group, Jennifer Harrow, Fotis E. Psomopoulos & Silvio C. E. Tosatto

Nature Methods (2021) | Cite this article

4927 Accesses | **73** Altmetric | Metrics

DOI: https://doi.org/10.1038/s41592-021-01205-4
preprint: https://arxiv.org/abs/2006.16189

CC BY SA

Data **O**ptimisation **M**odel **E**valuation

- Provenance
- Data splits
- Redundancy
- Availability

AN EXAMPLE DATA TABLE

Data	Provenance	*Protein Data Bank (PDB). X-ray structures missing residues. N_{pos} = 339,603 residues. N_{neg} = 6,168,717 residues. Previously used in (Walsh et al., Bioinformatics 2015) as an independent benchmark set.*
	Dataset splits	*training set: N/A* *$N_{pos,test}$ = 339,603 residues. $N_{neg,test}$ = 6,168,717 residues. No validation set. 5.22% positives on the test set.*
	Redundancy between data splits	*Not applicable.*
	Availability of data	*Yes, URL: http://protein.bio.unipd.it/mobidblite/.* *Free use license.*

Example paper used:
Marco Necci, et al, "MobiDB-lite: fast and highly specific consensus prediction of intrinsic disorder in proteins", Bioinformatics, 2017, https://doi.org/10.1093/bioinformatics/btx015

Data **O**ptimisation Model Evaluation

- Algorithm
- Meta-predictions
- Data encoding
- Parameters
- Features
- Fitting
- Availability

CC BY SA

AN EXAMPLE OPTIMIZATION TABLE

Optimization	Algorithm	*Majority-based consensus classification based on 8 primary ML methods and post-processing.*
	Meta-predictions	*Yes, predictor output is a binary prediction computed from the consensus of other methods; Independence of training sets of other methods with test set of meta-predictor was not tested since datasets from other methods were not available.*
	Data encoding	*Label-wise average of 8 binary predictions.*
	Parameters	*p = 3 (Consensus score threshold, expansion-erosion window, length threshold). No optimization.*
	Features	*Not applicable.*
	Fitting	*Single input ML methods are used with default parameters. Optimization is a simple majority.*
	Regularization	*No.*
	Availability of configuration	*Not applicable.*

Data Optimisation **Model** Evaluation

- Interpretability
- Execution time
- Software Availability

CC BY SA

AN EXAMPLE MODEL TABLE

Model	Interpretability	*Transparent, in so far as meta-prediction is concerned. Consensus and post processing over other methods predictions (which are mostly balck boxes). No attempt was made to make the meta-prediction a black box.*
	Output	*Classification, i.e. residues thought to be disordered.*
	Execution time	*ca. 1 second per representative on a desktop PC.*
	Availability of software	*Yes, URL: http://protein.bio.unipd.it/mobidblite/.* *Bespoke license free for academic use.*

Example paper used:
Marco Necci, et al, "MobiDB-lite: fast and highly specific consensus prediction of intrinsic disorder in proteins", Bioinformatics, 2017, https://doi.org/10.1093/bioinformatics/btx015

Data Optimisation Model **Evaluation**

- Evaluation
- Performance
- Comparison
- Confidence Availability

AN EXAMPLE EVALUATION TABLE

Evaluation	Evaluation method	*Independent dataset*
	Performance measures	*Balanced Accuracy, Precision, Sensitivity, Specificity, F1, MCC.*
	Comparison	*DisEmbl-465, DisEmbl-HL, ESpritz Disprot, ESpritz NMR, ESpritz Xray, Globplot, IUPred long, IUPred short, VSL2b. Chosen methods are the methods from which the meta prediction is obtained.*
	Confidence	*Not calculated.*
	Availability of evaluation	*No.*

Example paper used:
Marco Necci, et al, "MobiDB-lite: fast and highly specific consensus prediction of intrinsic disorder in proteins", Bioinformatics, 2017, https://doi.org/10.1093/bioinformatics/btx015

ROADMAP FOR THE ELIXIR MACHINE LEARNING FOCUS GROUP

Short term activities

Long term activities

Activity 1

Support the **automation** of the DOME recommendations, assisting researchers to effectively report on their work. This will be complemented by the design of a **controlled vocabulary** relevant to each DOME field.

infrastructure activity

Activity 2

Connect and review the **gold-standard dataset definitions** in relevant groups (including Health Data FG, Cancer Data FG, BM1G, text mining data) and engaging with communities.

standards definition activity

Activity 3

Review **synthetic data needs** w.r.t ML such as existing synthetic datasets, generation tools, federated generation etc, and coordinate with other groups that are doing related efforts.

discussion starting in this area

Governance

Governance structure around producing and maintaining ML standards in life sciences, with ELIXIR at the centre, and involving all relevant stakeholders (CLAIRE/Pistoia Alliance/ Kipoi/RDA/ReSA).

Implementation

Establishing infrastructure services around ML, as a horizontal, cross-platform effort.

Designing and implementing a Machine Actionable version of the DOME recommendations.

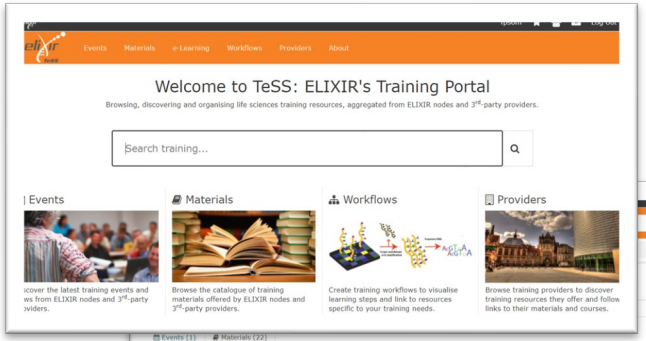
CLOSING NOTE
TRAINING AS A FACILITATOR

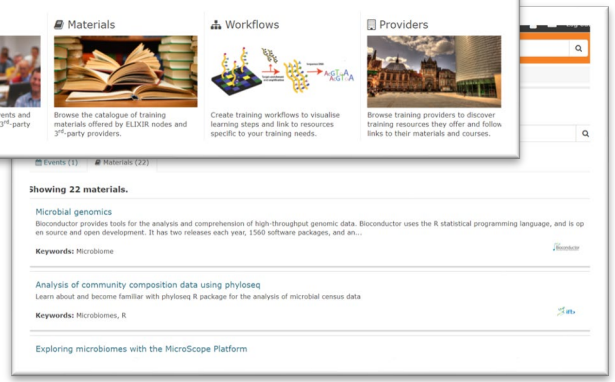
"A man is only as good as his tools"
Emmert Wolf

Adopting and using a standard is, by itself, a challenge.

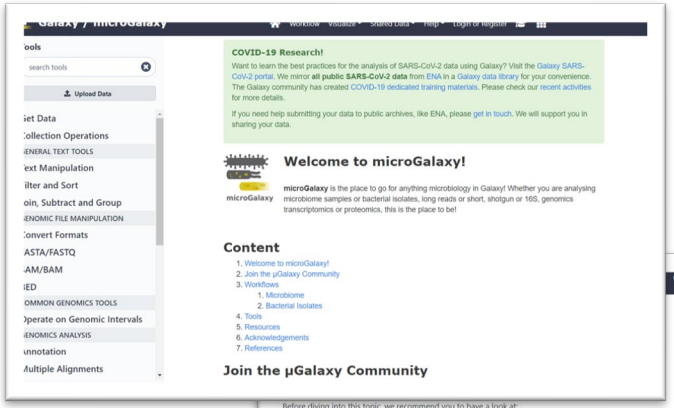
Several organizations that offer training in Life Sciences, and could act as a catalyst:

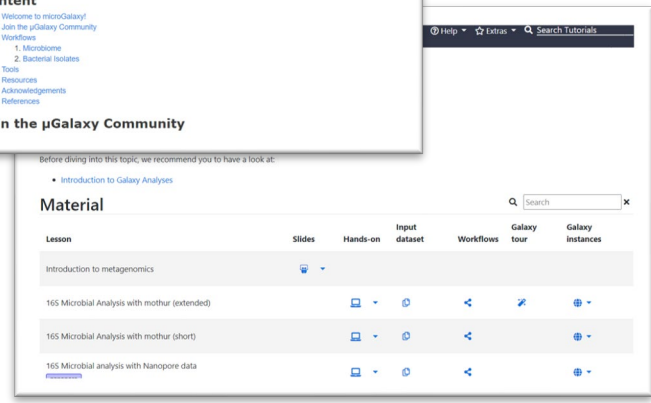
https://elixir-europe.org/





https://tess.elixir-uk.org/





https://training.galaxyproject.org/training-material/

https://gallantries.github.io/

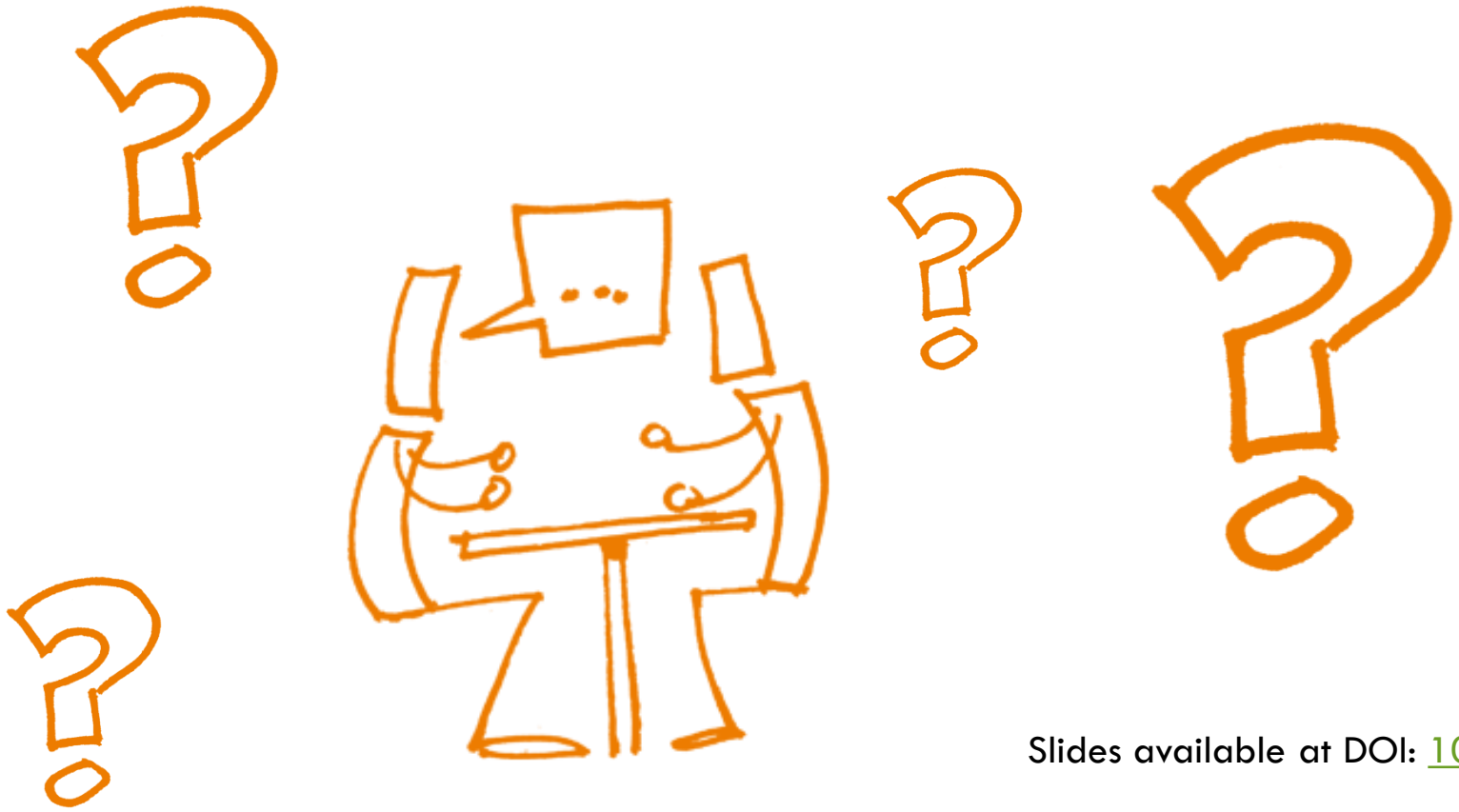
ACKNOWLEDGEMENTS

- Jen Harrow (ELIXIR-Hub)
- Silvio Tosatto (Un. of Padova, ELIXIR-IT)
- The ELIXIR Machine Learning Focus Group

UNIVERSITÀ DEGLI STUDI DI PADOVA

CC BY SA

THANK YOU FOR YOUR ATTENTION!



Slides available at DOI: 10.5281/zenodo.5566576

@fopsom

CC BY SA