Deep Learning for AI: turning the GPS on

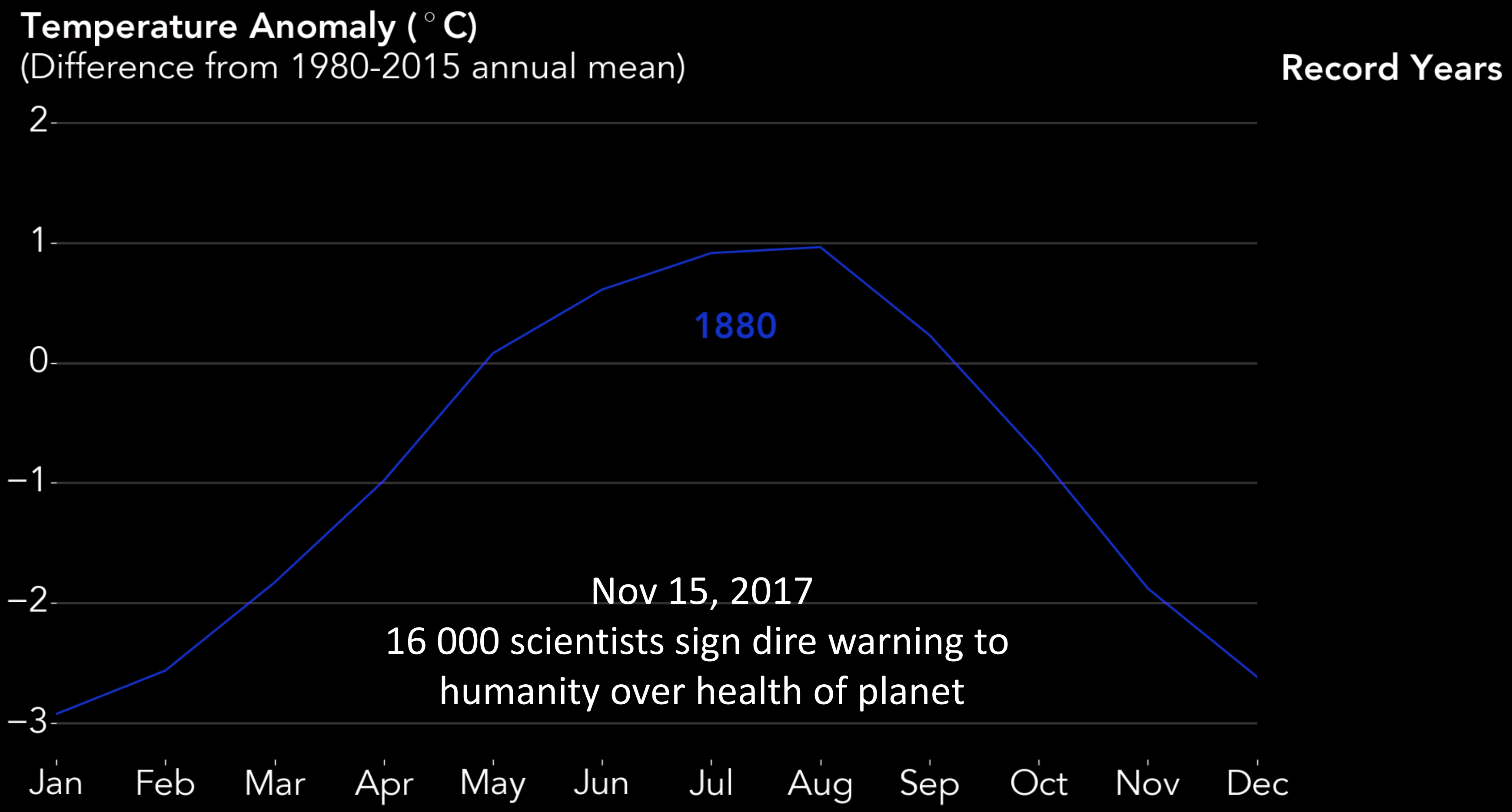
Cesare Furlanello

FBK // DATASCIENCE // MPBA

@furlanello

MPBA
PREDICTIVE MODELS
FOR BIOMEDICINE & ENVIRONMENT

FONDAZIONE
BRUNO KESSLER



Nov 15, 2017
16 000 scientists sign dire warning to humanity over health of planet

changing the world

@furlanello

Credits: NASA-NOAA

mpbalab.fbk.eu

Smart Cities & Communities, Health & Well Being, Future Media, Machine Translation, Smart Digital Industry, Data Science, CyberSecurity

FBK for Artificial Intelligence – 2018

REVOLUTIONARY CHANGE OF LANDSCAPE IN RESEARCH & INNOVATION

Private traits and attributes are predictable from digital records of human behavior

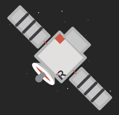
IOT, Industry, Retail, Finance, Healthcare, Agricolture already changed

Machine Learning as commodity (0.9 $/h)

0.3-0.5% of the population knows how to code

WHERE TO START ?

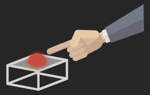
PRODUCTS


OBSERVATION

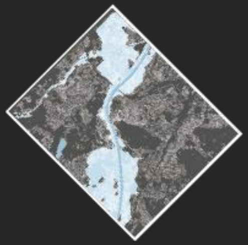
FORECAST

INTEGRATION OF EXISTING SYSTEMS

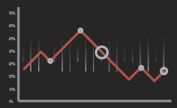
UNMANNED AI SOLUTIONS


DECISION SUPPORT SYSTEMS

BIG / DATA

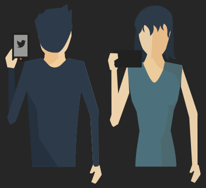

DATA LANDSCAPE

ORGANIZATION PROCESS


HISTORICAL DATA

SOCIAL MEDIA DATA STREAMS

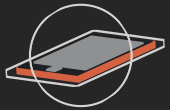
HUMAN RESOURCES

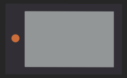

AI / ML PRACTITIONERS

?

C. Furlanello – MPBA Nov2017

TECHNOLOGIES


SMARTPHONE APP


WEB INTERFACES

WEARABLES

Credits:
I-REACT H2020

UNDER THE HOOD

"It's ML, not magic"

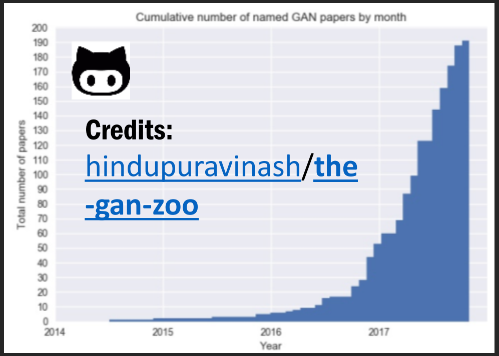
Credits: @smerity

- The Data Science stack (Python, R) and resources (arXiv.org)
- Keras with TensorFlow backend
- PyTorch
- Fast, well tuned baselines
- Model Selection: human intuition in Deep Learning is bad
- Ability to accurately measure progress over time

OK, but … do I need a GPU Armada?

FLOOD & FIRE

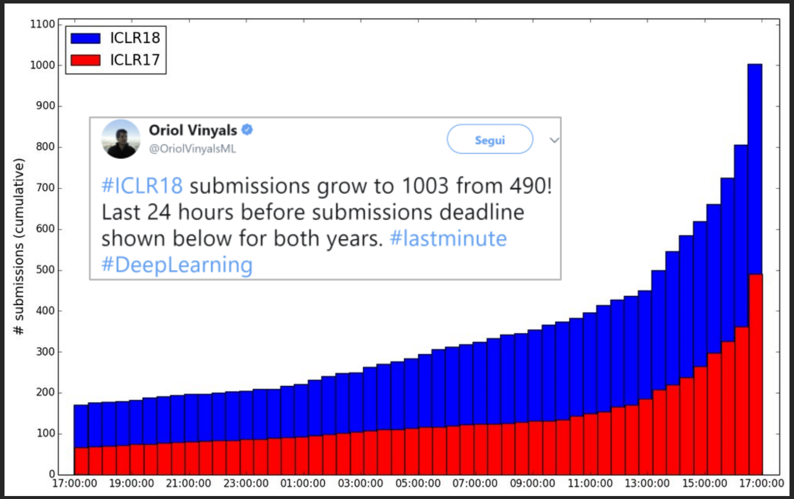
Generative Adversarial Networks (GANs) new models only



Credits: hindupuravinash/the-gan-zoo

unprecedented and growing cost of keeping track of new DL ideas

- **Karpathy's http://www.arxiv-sanity.com/**
- **http://search.iclr2018.smerity.com/: hits RNN: 324, CNN: 455, GAN: 666**



Oriol Vinyals
@OriolVinyalsML

#ICLR18 submissions grow to 1003 from 490! Last 24 hours before submissions deadline shown below for both years. #lastminute #DeepLearning

MPBA LAB

Data Science Lab (25 pe) : Maths, CS, BioEngineering, Physics, ...

Unifarm’s robotic warehouse

- Machine Learning: systems that can learn from examples and predict over novel data
- Applied: Predictive Biomarkers, agri-tech; environmental risk, IoT, personal sensors

2 ½ STARTUPS

- MPA Solutions: geospatial analytics
- Motorialab: sport analytics ➔ insurtech
- Multipl.AI: DL for Precision Medicine

ML Projects with industry and retail

EIT Digital: Wearable Analytics, Smart Retail, Fraud profiling

MPBA LAB

- ML as infrastructure: Big Data Analytics, Networks, GIS, bio-informatics, Deep Learning
- Fast deployement of dashboard analytics and microservices (CPU/GPU) in cloud

Forecasting EDM Ticket Sales with ML & Networks

EXCELLENCE IN RESEARCH

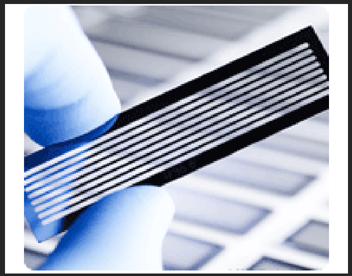
- Precision Medicine, with FDA, OPBG, Riken; HEP with Liverpool

Dashboards as enablers of “actionable” analytics: explore patterns, interact with models for simulation

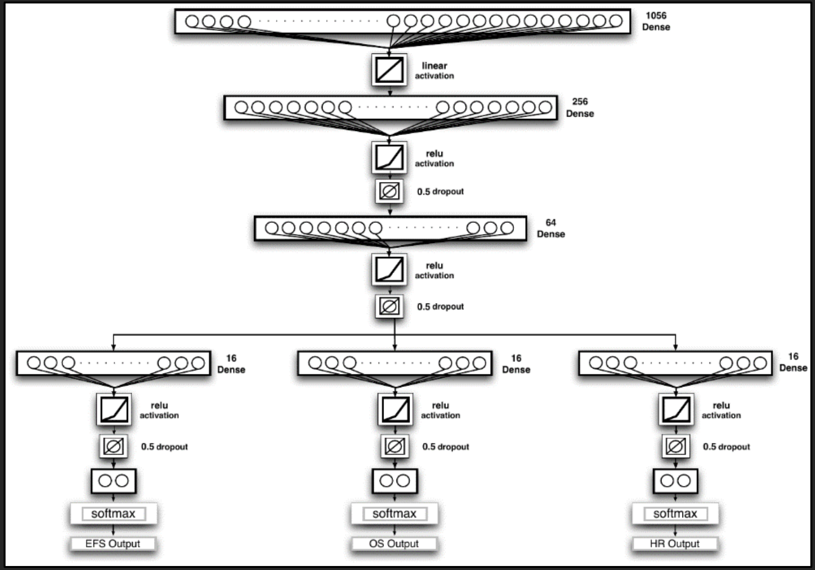
CHALLENGES FOR DEEP LEARNING

- Data integration & reproducibility in diagnostics and pharmacogenomics
- Deep Learning as accelerator of new ideas (e.g. embeddings, GANs)

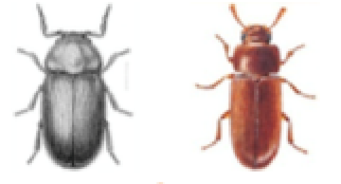
MPBA: DEEP LEARNING FOR MASSIVE DATA



Multi-Objective Deep Learning on Massive Sequencing Data

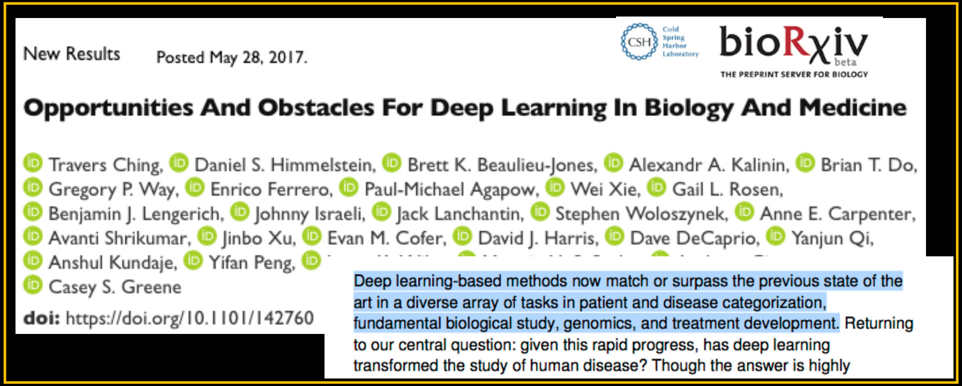


A. SEQC 2017: QC of massive NGS data for Precision Medicine



B. Food safety

Deep Learning & Omics



New Results Posted May 28, 2017.

CSH Cold Spring Harbor Laboratory
bioRχiv beta
THE PREPRINT SERVER FOR BIOLOGY

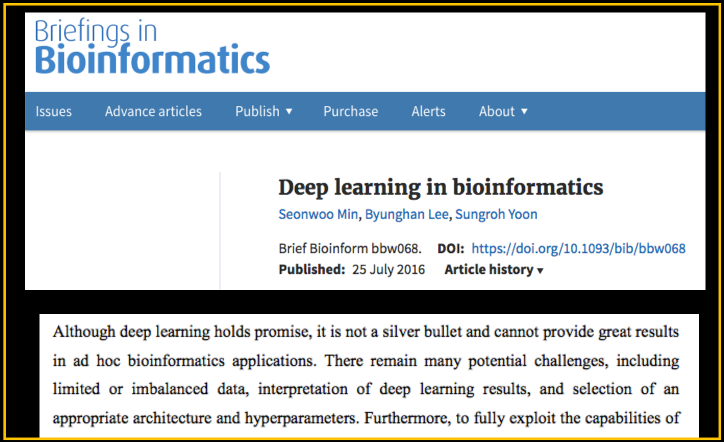
Opportunities And Obstacles For Deep Learning In Biology And Medicine

Travers Ching, Daniel S. Himmelstein, Brett K. Beaulieu-Jones, Alexandr A. Kalinin, Brian T. Do, Gregory P. Way, Enrico Ferrero, Paul-Michael Agapow, Wei Xie, Gail L. Rosen, Benjamin J. Lengerich, Johnny Israeli, Jack Lanchantin, Stephen Woloszynek, Anne E. Carpenter, Avanti Shrikumar, Jinbo Xu, Evan M. Cofer, David J. Harris, Dave DeCaprio, Yanjun Qi, Anshul Kundaje, Yifan Peng, Casey S. Greene

doi: https://doi.org/10.1101/142760

Deep learning-based methods now match or surpass the previous state of the art in a diverse array of tasks in patient and disease categorization, fundamental biological study, genomics, and treatment development. Returning to our central question: given this rapid progress, has deep learning transformed the study of human disease? Though the answer is highly

🟢



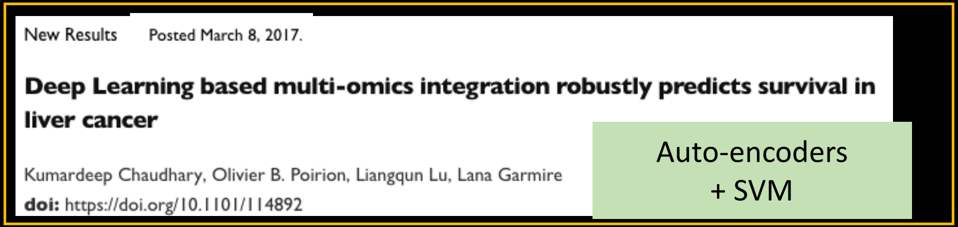
Deep learning in bioinformatics

Seonwoo Min, Byunghan Lee, Sungroh Yoon

Brief Bioinform bbw068. DOI: https://doi.org/10.1093/bib/bbw068
Published: 25 July 2016 Article history ▾

Although deep learning holds promise, it is not a silver bullet and cannot provide great results in ad hoc bioinformatics applications. There remain many potential challenges, including limited or imbalanced data, interpretation of deep learning results, and selection of an appropriate architecture and hyperparameters. Furthermore, to fully exploit the capabilities of

🟡



New Results Posted March 8, 2017.

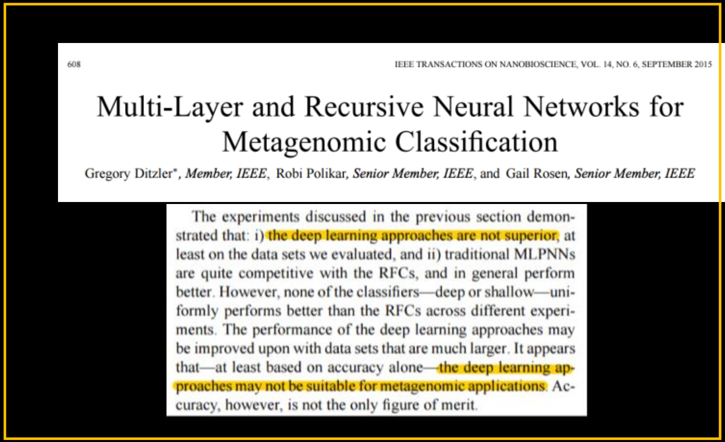
Deep Learning based multi-omics integration robustly predicts survival in liver cancer

Kumardeep Chaudhary, Olivier B. Poirion, Liangqun Lu, Lana Garmire

doi: https://doi.org/10.1101/114892

Auto-encoders
+ SVM

🟢



608 IEEE TRANSACTIONS ON NANOBIOSCIENCE, VOL. 14, NO. 6, SEPTEMBER 2015

Multi-Layer and Recursive Neural Networks for Metagenomic Classification

Gregory Ditzler*, *Member, IEEE*, Robi Polikar, *Senior Member, IEEE*, and Gail Rosen, *Senior Member, IEEE*

The experiments discussed in the previous section demonstrated that: i) the deep learning approaches are not superior, at least on the data sets we evaluated, and ii) traditional MLPNNs are quite competitive with the RFCs, and in general perform better. However, none of the classifiers—deep or shallow—uniformly performs better than the RFCs across different experiments. The performance of the deep learning approaches may be improved upon with data sets that are much larger. It appears that—at least based on accuracy alone—the deep learning approaches may not be suitable for metagenomic applications. Accuracy, however, is not the only figure of merit.

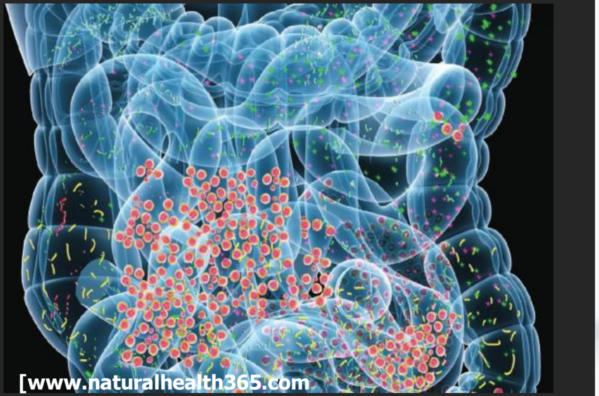
🔴

CHANGING THE COURSE OF GENOMIC MEDICINE

Today, machine learning and experimental biology are advancing at an exponential pace. Deep Genomics is where these disciplines meet. Our systems predict the molecular effec variation, opening a new and exciting path to discovery for disease diagnostics an

🟢

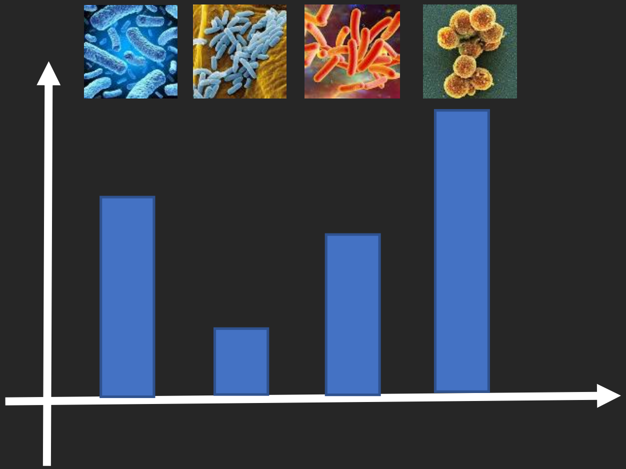
Metagenomics for Gut Inflammatory Disease



[www.naturalhealth365.com
]

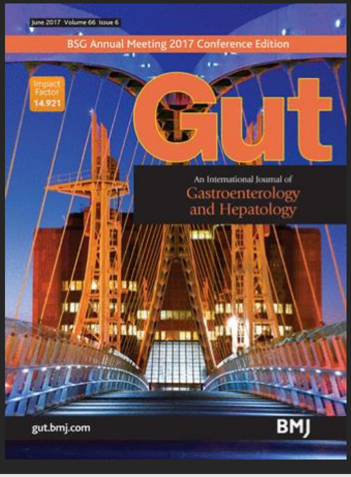
Gut microbiome

Next Generation Sequencing



Bacterial abundances

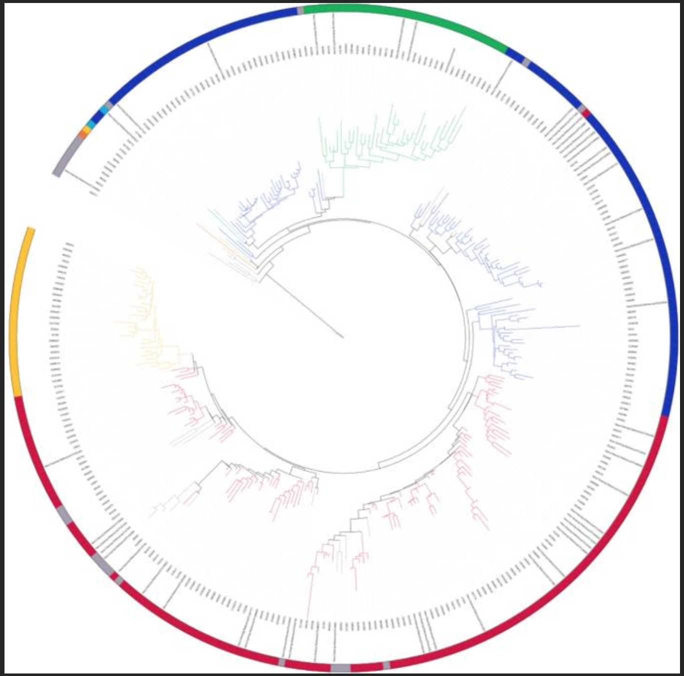
Bioinformatics pipeline


June 2017 Volume 66 Issue 6
BSG Annual Meeting 2017 Conference Edition
Impact Factor 14.921
Gut
An International Journal of
Gastroenterology and Hepatology
gut.bmj.com
BMJ

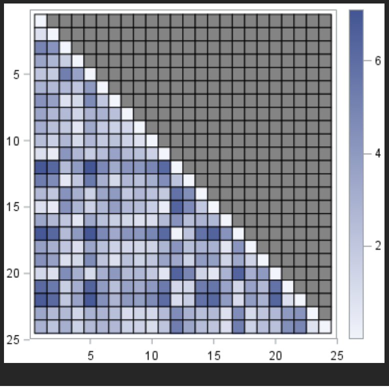
Inflammatory Bowel Disease, with H Sokol

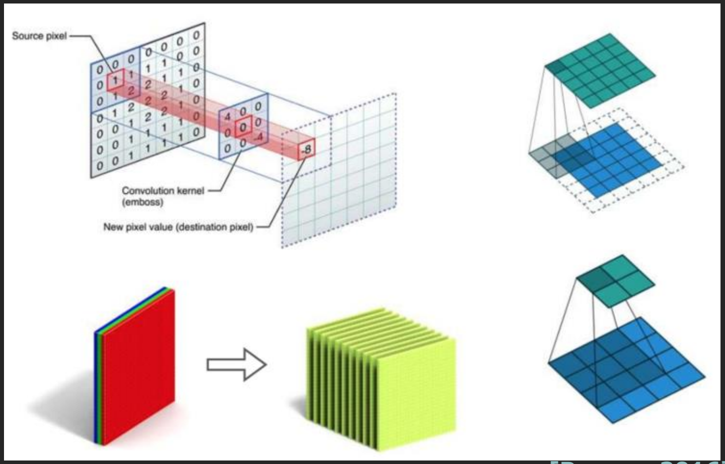
A CNN architecture for Metagenomics

Phylogenetic tree



Distance matrix





[Perone, 2016]

Convolutional Neural Networks

Defining a tree distance from “Patristic”

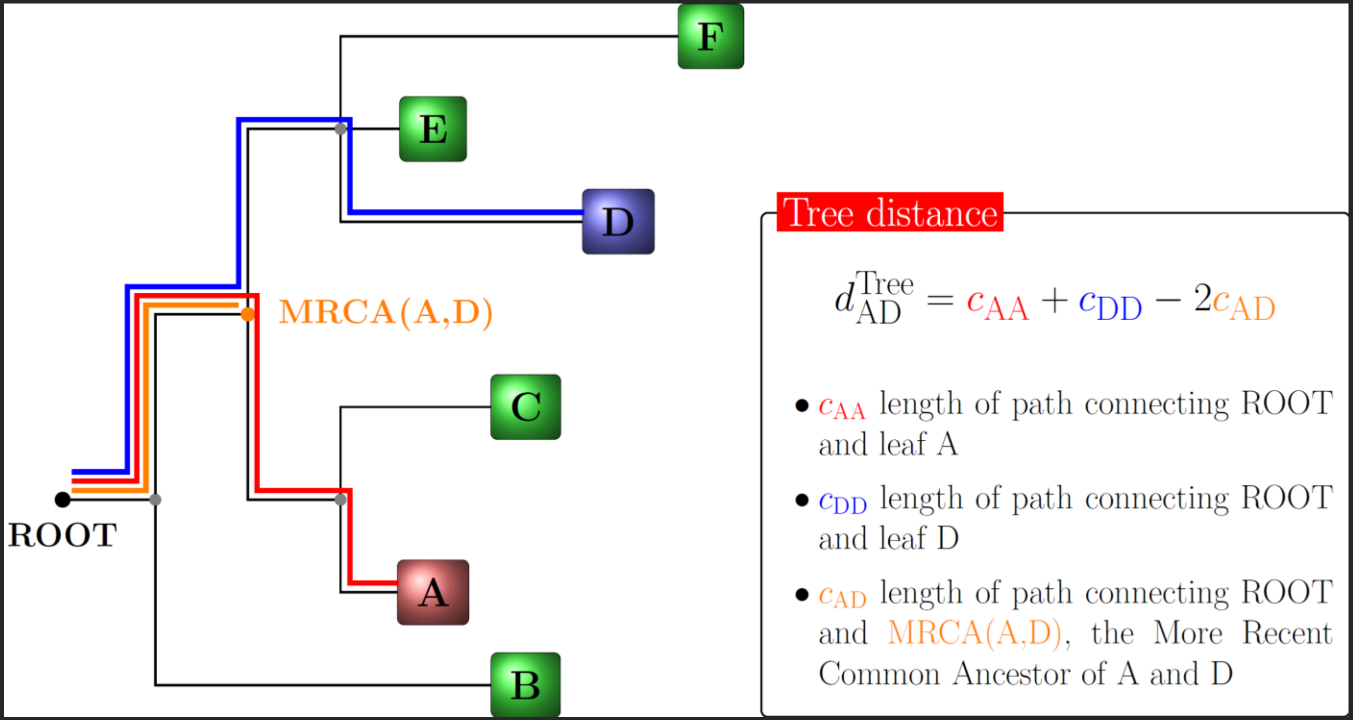
Key ingredient:

Concept nearest neighbours

d(,)

Patristic distance:

Σ(length of all branches) connecting two species

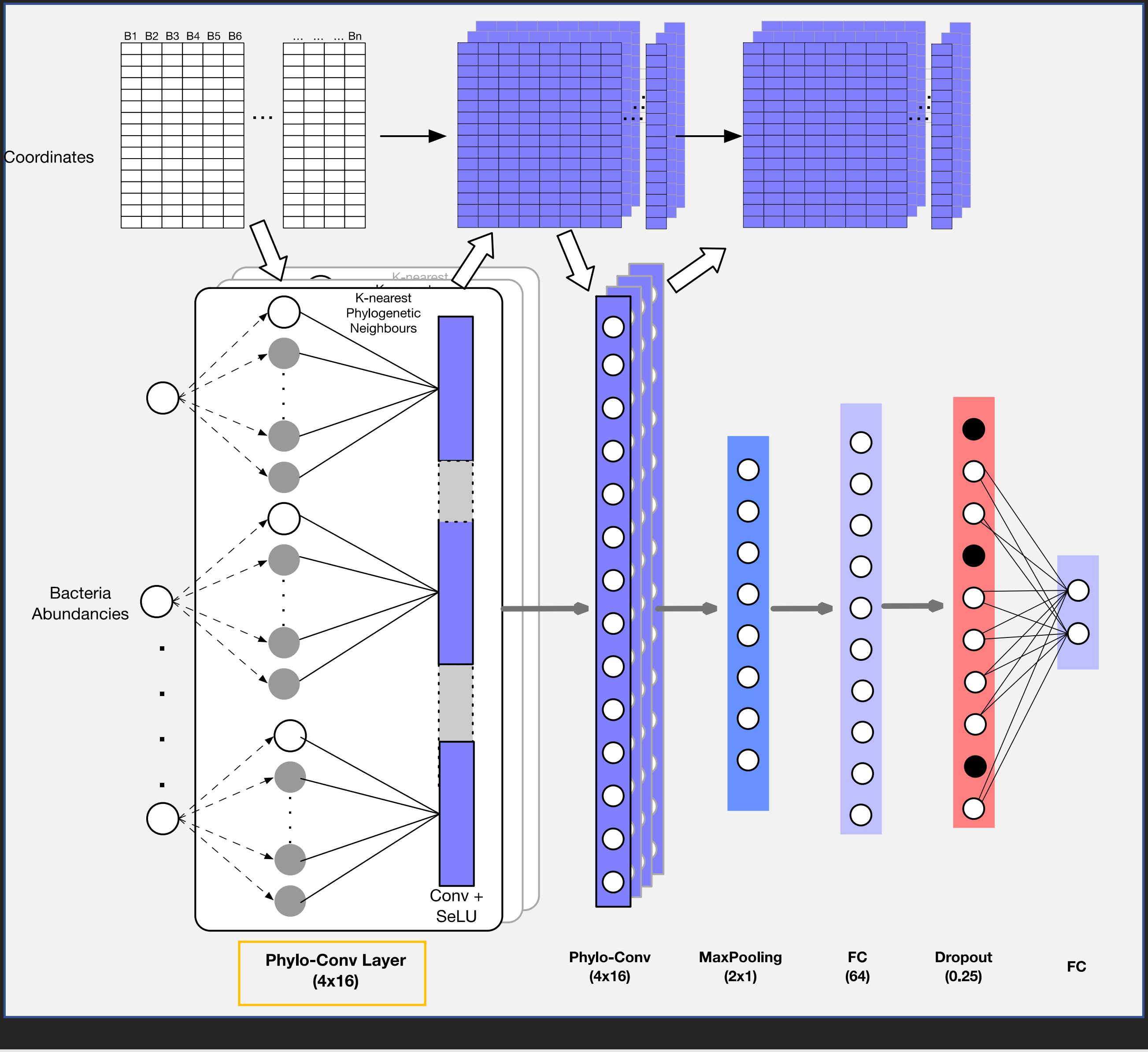


Tree distance

$$d^{\mathrm{Tree}}_{\mathrm{AD}} = c_{\mathrm{AA}} + c_{\mathrm{DD}} - 2c_{\mathrm{AD}}$$

- c_{AA} length of path connecting ROOT and leaf A
- c_{DD} length of path connecting ROOT and leaf D
- c_{AD} length of path connecting ROOT and MRCA(A,D), the More Recent Common Ancestor of A and D

Ph-CNN



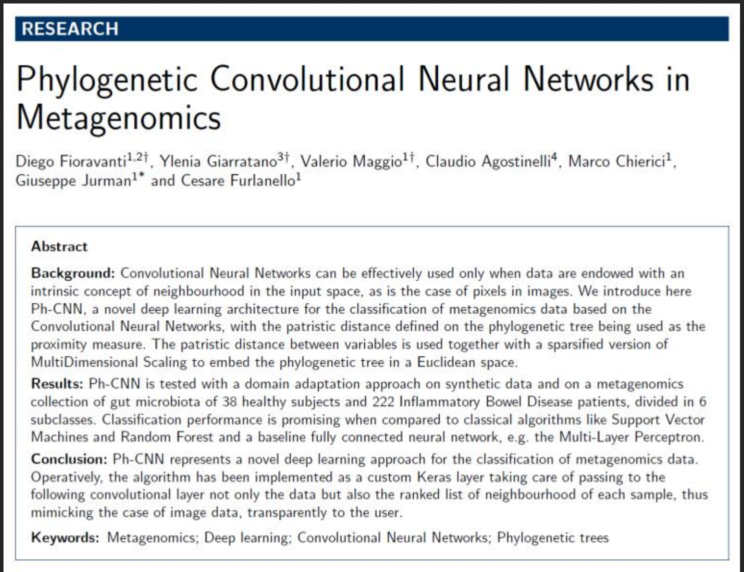
python™

K + TensorFlow

IMPLEMENTATION

- A new Keras layer
- TensorFlow
- GPU Cloud
- ADAM optimizer

GENERALIZATION TO OmicsCNN



RESEARCH

Phylogenetic Convolutional Neural Networks in Metagenomics

Diego Fioravanti[1,2†], Ylenia Giarratano[3†], Valerio Maggio[1†], Claudio Agostinelli[4], Marco Chierici[1], Giuseppe Jurman[1*] and Cesare Furlanello[1]

Abstract

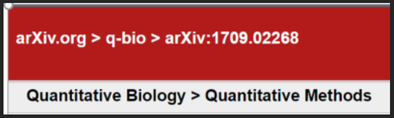
Background: Convolutional Neural Networks can be effectively used only when data are endowed with an intrinsic concept of neighbourhood in the input space, as is the case of pixels in images. We introduce here Ph-CNN, a novel deep learning architecture for the classification of metagenomics data based on the Convolutional Neural Networks, with the patristic distance defined on the phylogenetic tree being used as the proximity measure. The patristic distance between variables is used together with a sparsified version of MultiDimensional Scaling to embed the phylogenetic tree in a Euclidean space.

Results: Ph-CNN is tested with a domain adaptation approach on synthetic data and on a metagenomics collection of gut microbiota of 38 healthy subjects and 222 Inflammatory Bowel Disease patients, divided in 6 subclasses. Classification performance is promising when compared to classical algorithms like Support Vector Machines and Random Forest and a baseline fully connected neural network, e.g. the Multi-Layer Perceptron.

Conclusion: Ph-CNN represents a novel deep learning approach for the classification of metagenomics data. Operatively, the algorithm has been implemented as a custom Keras layer taking care of passing to the following convolutional layer not only the data but also the ranked list of neighbourhood of each sample, thus mimicking the case of image data, transparently to the user.

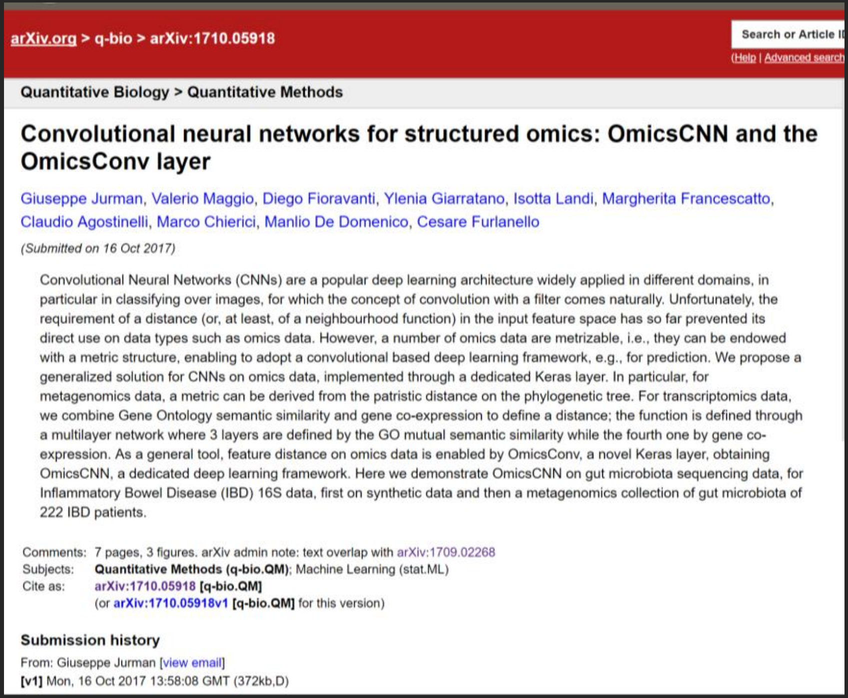
Keywords: Metagenomics; Deep learning; Convolutional Neural Networks; Phylogenetic trees

Fioravanti et al 2017, BMC Bioinformatics, in press



arXiv.org > q-bio > arXiv:1709.02268

Quantitative Biology > Quantitative Methods



arXiv.org > q-bio > arXiv:1710.05918

Quantitative Biology > Quantitative Methods

Convolutional neural networks for structured omics: OmicsCNN and the OmicsConv layer

Giuseppe Jurman, Valerio Maggio, Diego Fioravanti, Ylenia Giarratano, Isotta Landi, Margherita Francescatto, Claudio Agostinelli, Marco Chierici, Manlio De Domenico, Cesare Furlanello

(Submitted on 16 Oct 2017)

Convolutional Neural Networks (CNNs) are a popular deep learning architecture widely applied in different domains, in particular in classifying over images, for which the concept of convolution with a filter comes naturally. Unfortunately, the requirement of a distance (or, at least, of a neighbourhood function) in the input feature space has so far prevented its direct use on data types such as omics data. However, a number of omics data are metrizable, i.e., they can be endowed with a metric structure, enabling to adopt a convolutional based deep learning framework, e.g., for prediction. We propose a generalized solution for CNNs on omics data, implemented through a dedicated Keras layer. In particular, for metagenomics data, a metric can be derived from the patristic distance on the phylogenetic tree. For transcriptomics data, we combine Gene Ontology semantic similarity and gene co-expression to define a distance; the function is defined through a multilayer network where 3 layers are defined by the GO mutual semantic similarity while the fourth one by gene co-expression. As a general tool, feature distance on omics data is enabled by OmicsConv, a novel Keras layer, obtaining OmicsCNN, a dedicated deep learning framework. Here we demonstrate OmicsCNN on gut microbiota sequencing data, for Inflammatory Bowel Disease (IBD) 16S data, first on synthetic data and then a metagenomics collection of gut microbiota of 222 IBD patients.

Comments:	7 pages, 3 figures. arXiv admin note: text overlap with arXiv:1709.02268
Subjects:	**Quantitative Methods (q-bio.QM)**; Machine Learning (stat.ML)
Cite as:	arXiv:1710.05918 **[q-bio.QM]**
	(or **arXiv:1710.05918v1 [q-bio.QM]** for this version)

Submission history

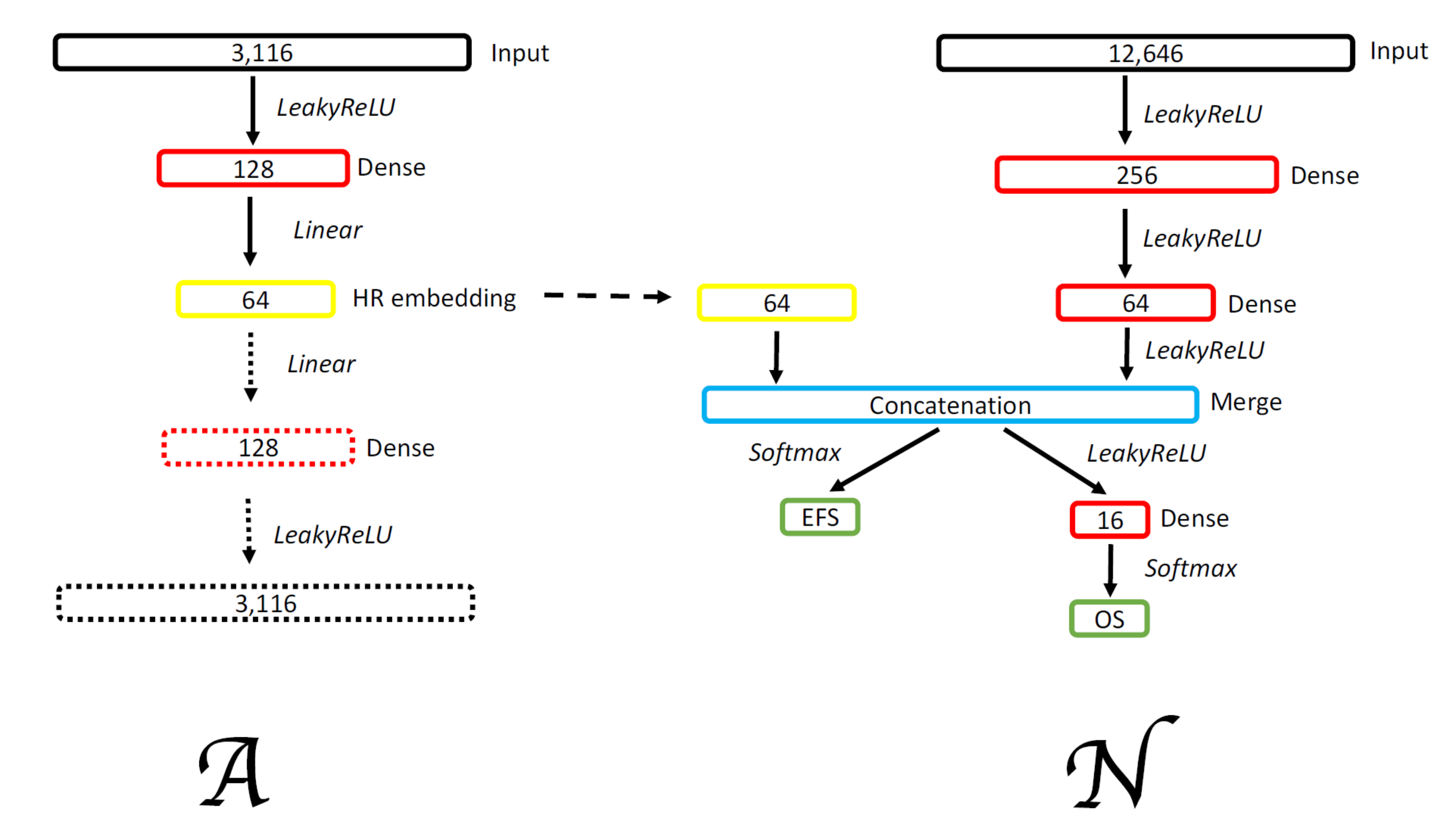
From: Giuseppe Jurman [view email]
[v1] Mon, 16 Oct 2017 13:58:08 GMT (372kb,D)

Jurman et al 2017 “Convolutional neural networks for structured omics: OmicsCNN and the OmicsConv layer

Machine Learning in Computational Biology
NIPS 2017 (9 Dec)

arXiv:1710.05918

Deep Learning for Diagnosis and Prognosis of Pediatric Cancer (500 RNA-Seq)



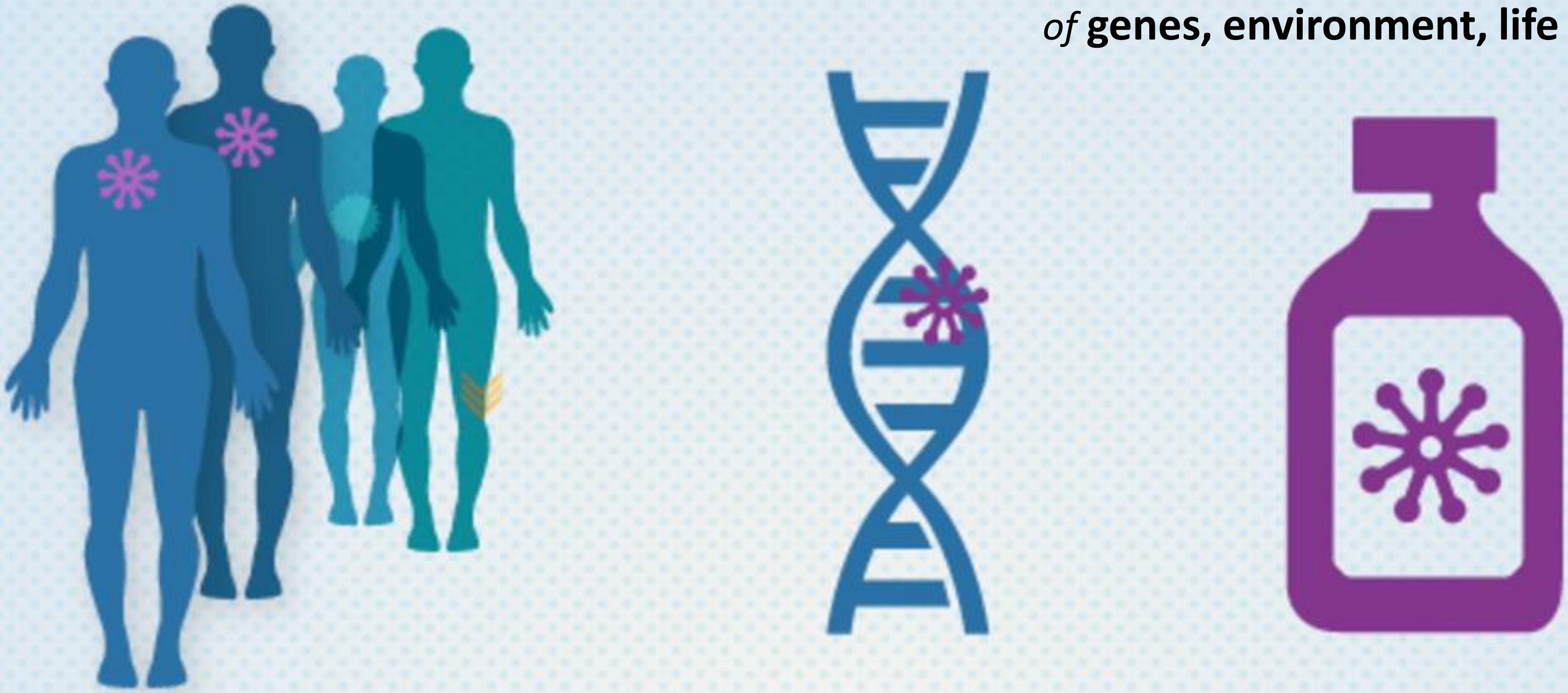
Maggio et al 2017 " MULTIOBJECTIVE Deep Learning Approach for predictive classification in Neuroblastoma"
ML4H: Machine Learning for Health Workshop NIPS 2017 (Dec 8, 2017)

PRECISION MEDICINE

A unifying framework

- *“Treating the **right patient** for the **right disease** at the **right time** with the **right amount of (the right) drug”***

In Diagnostics, treatment and prevention, *we shall systematically include the* ***individual variability*** *of* **genes, environment, life style**

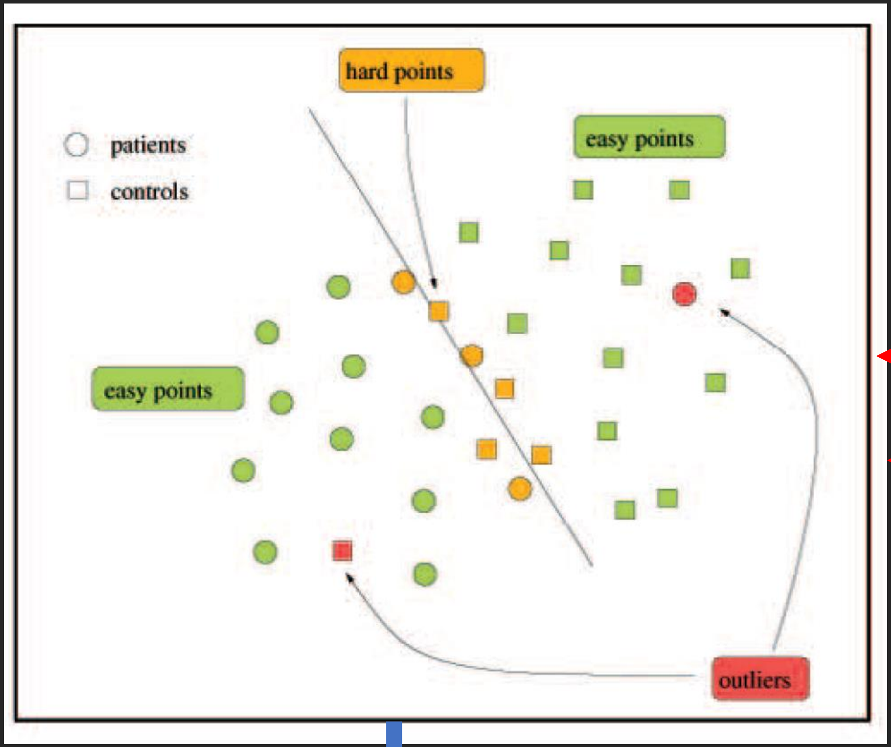


The Precision Medicine Initiative (USA 2015) is the vision that redefines healthcare and Pharma R&D

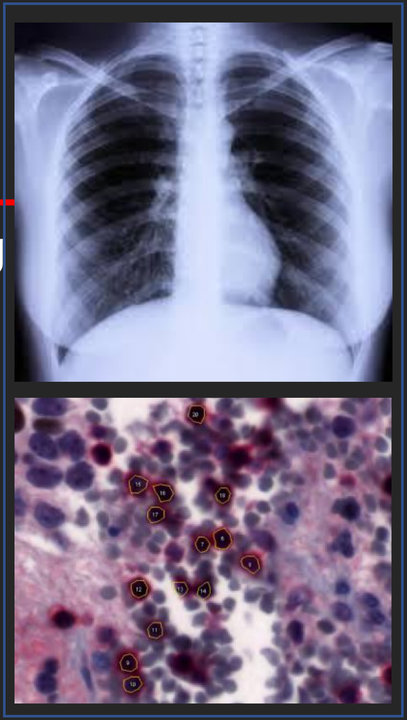
C. Furlanello – MPBA Nov 2017

C. Furlanello – MPBA Nov 2017

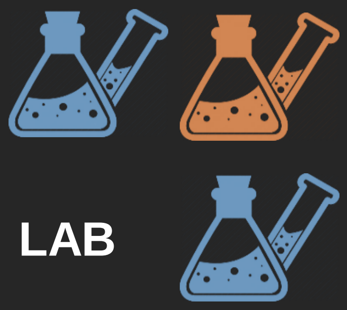
MACHINE LEARNING IN BIOMEDICINE



Next Generation Sequencing



Bioimaging

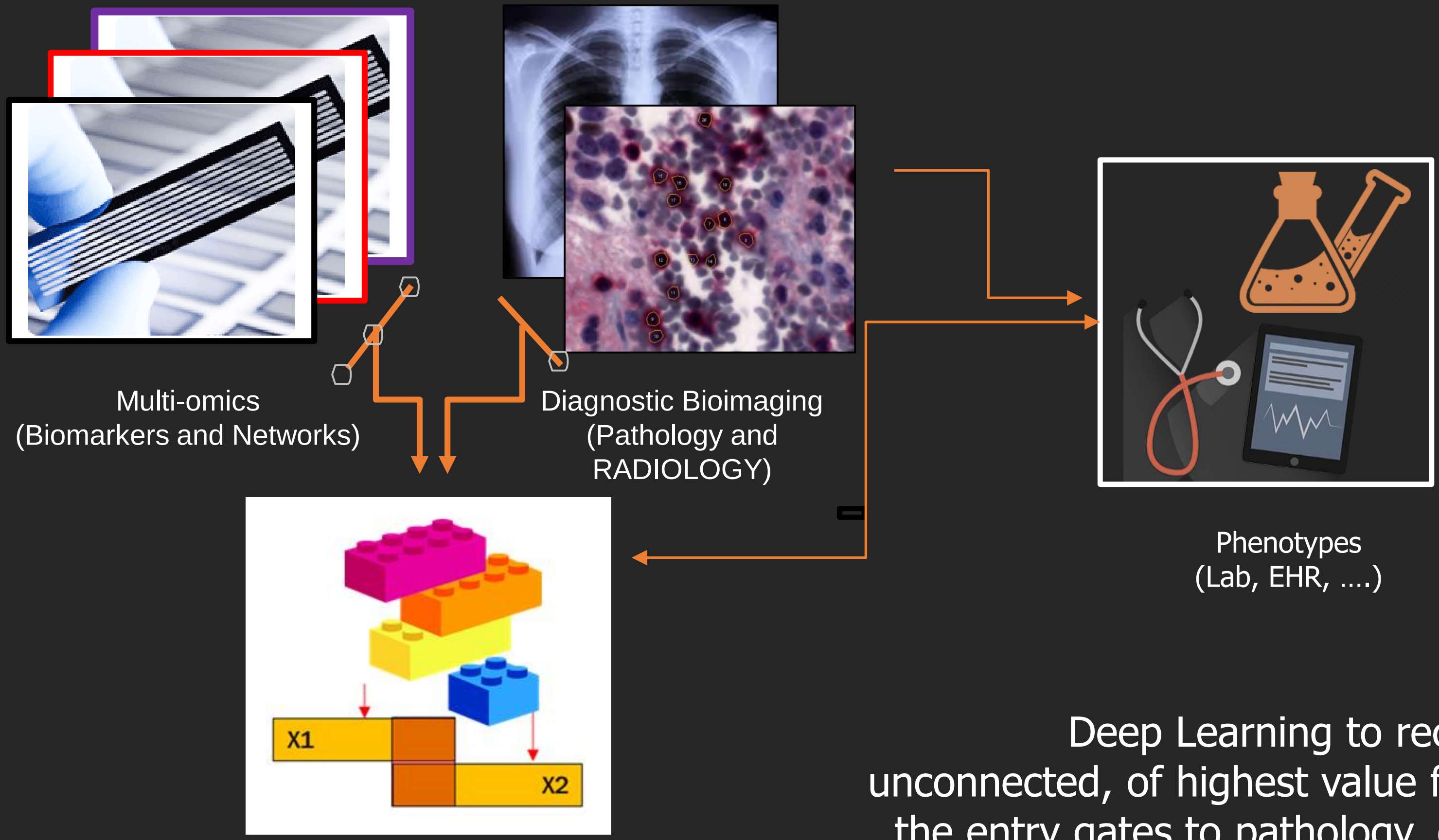


LAB

Outcome
(complex phenotype)

CHALLENGE

Genomic Medicine R&D is a patchwork in terms of data, technologies and models hard to integrate. We are studying a Deep Learning solution to identify a common share space for health trajectories



Multi-omics
(Biomarkers and Networks)

Diagnostic Bioimaging
(Pathology and
RADIOLOGY)

Phenotypes
(Lab, EHR, ...)

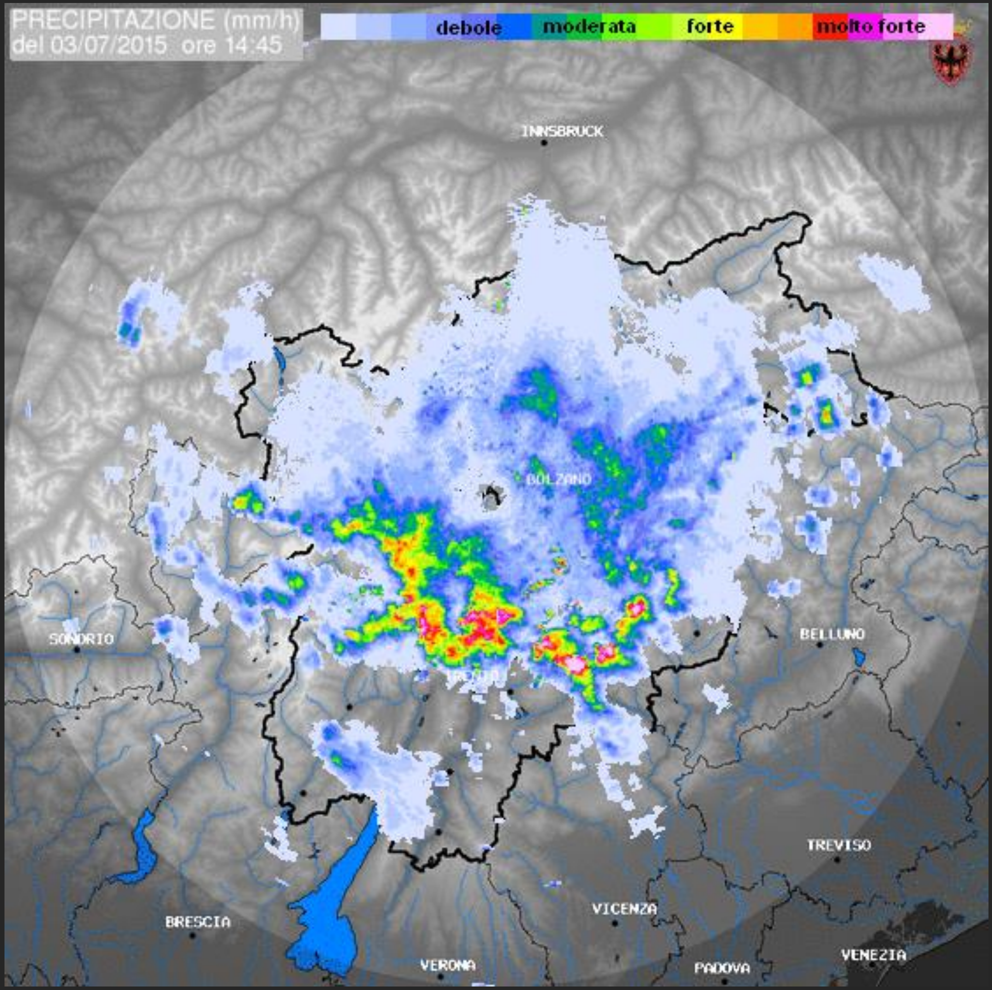
Natural integration in
the latent space

Multiple Endpoints (Diagnostics, prognostic, actionable)

Modulated by **Complex Clinical Phenotypes**

Deep Learning to recover data still unconnected, of highest value for reconstruct the entry gates to pathology, or to intercept their similarity, or transferable therapy between diseases.

FROM HEALTH TO ENVIRONMENT, AND BACK



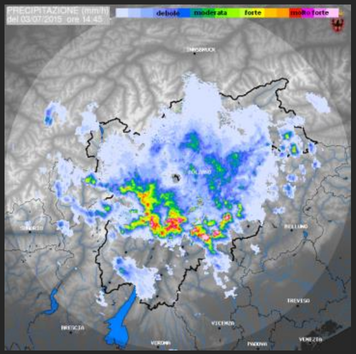
Spatio-temporal NOWCASTING (Conv-LSTM)

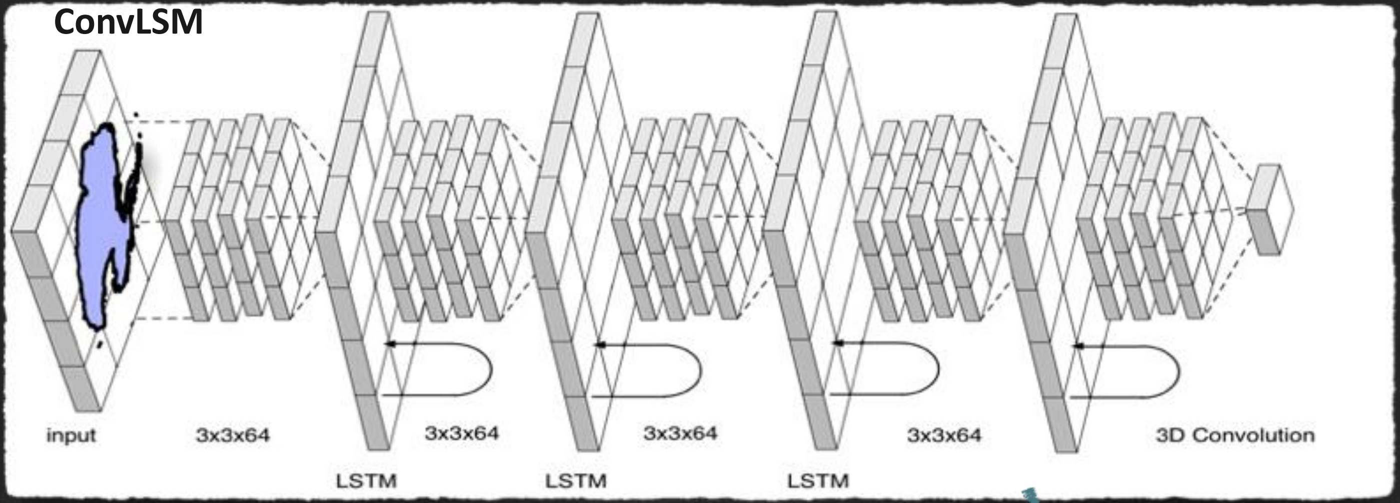
C. Furlanello – MPBA Nov2017

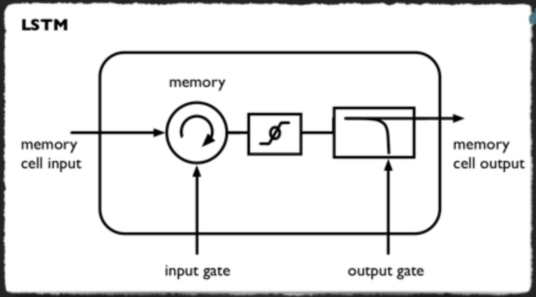
NOWCASTING

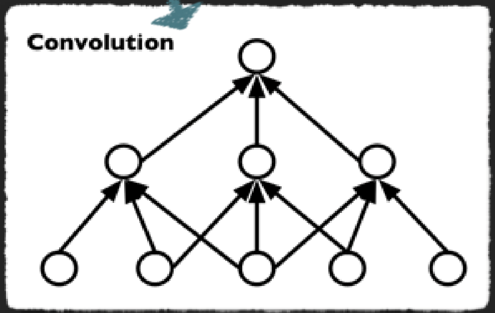
Rain & lightning nowcasting (5’ - 75’)

Short-time radar prediction: target for **deep learning**







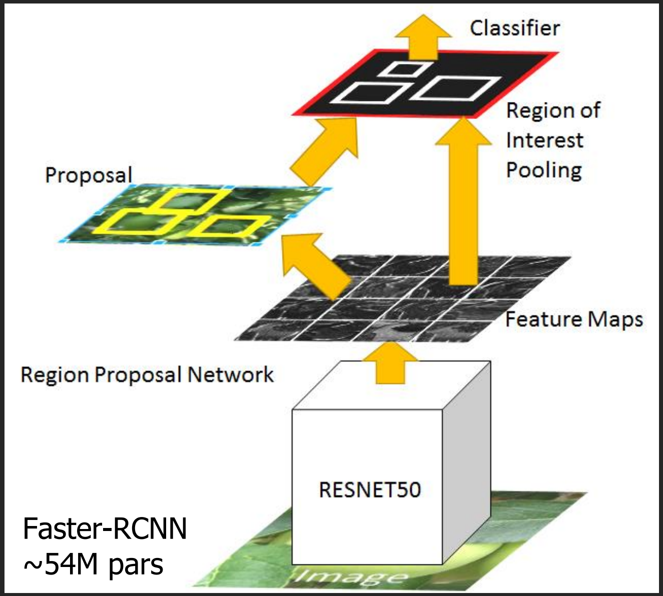


Deep Learning in AgriTech

MPBA
PREDICTIVE MODELS
FOR BIOMEDICINE & ENVIRONMENT

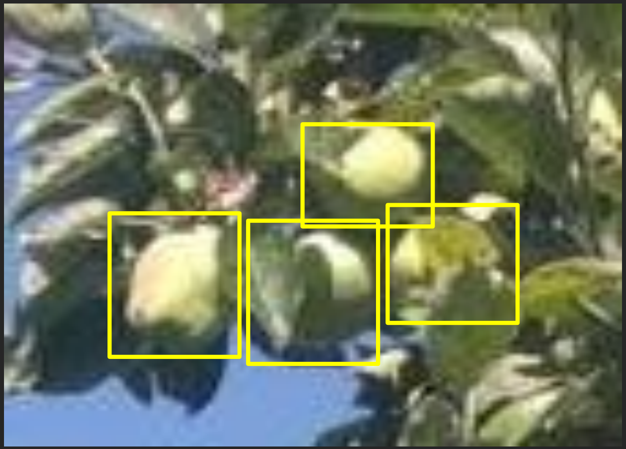
Expert estimating damage

- Time-consuming methods based on limited data
- High variance



Faster-RCNN
~54M pars

1. Yield quantification
2. Assessment of damage and risk
3. Quality Control



Tested July 2017, Val di Non, Trentino

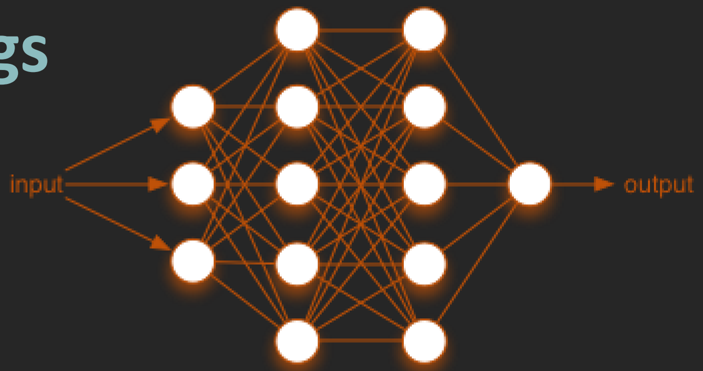
FONDAZIONE
BRUNO KESSLER

Deep Learning as a Service

Deep Learning for retail

Embeddings
(categorical)

Sales
(Numeric)



Target: daily sellout forecast (60 days)

What: Deep Learning pipeline for POS sales

First large scale application: pharmaceutical retail (UNIFARM),

- **315K unique products, 458 POS** over **1642 days** (~4 years),
- Total sales of **€ 2.4 Billions**
- The module is **accurate** (improves over XGBoost): **16% RMSPE**

eit Digital

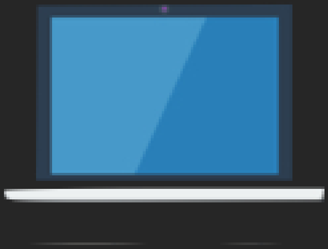
Partners: FBK, REPLY, Tim, BT, DFKI Saarbrucken …

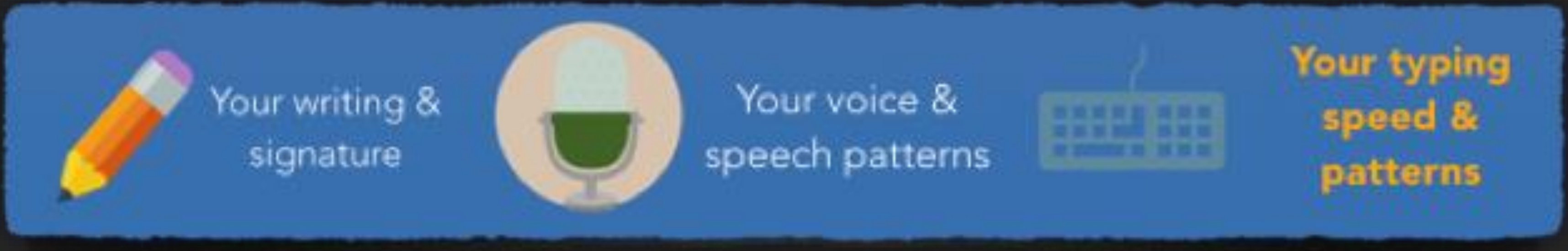
DL User profiling (fintech security)

BIOLOGICAL
BIOMETRICS
Your iris & retina
Your fingerprints
Your DNA

Keystroke dynamics: how a user types

- **Monitor keyboard inputs (KHz),**
- **Process typing features**
- **Defines a pattern for future comparison**





C. Furlanello – MPBA Nov2017

DWELL TIME

The time between pressure and release of a key

GAP TIME

The lapse of time between striking one key and the next

Your typing speed & patterns

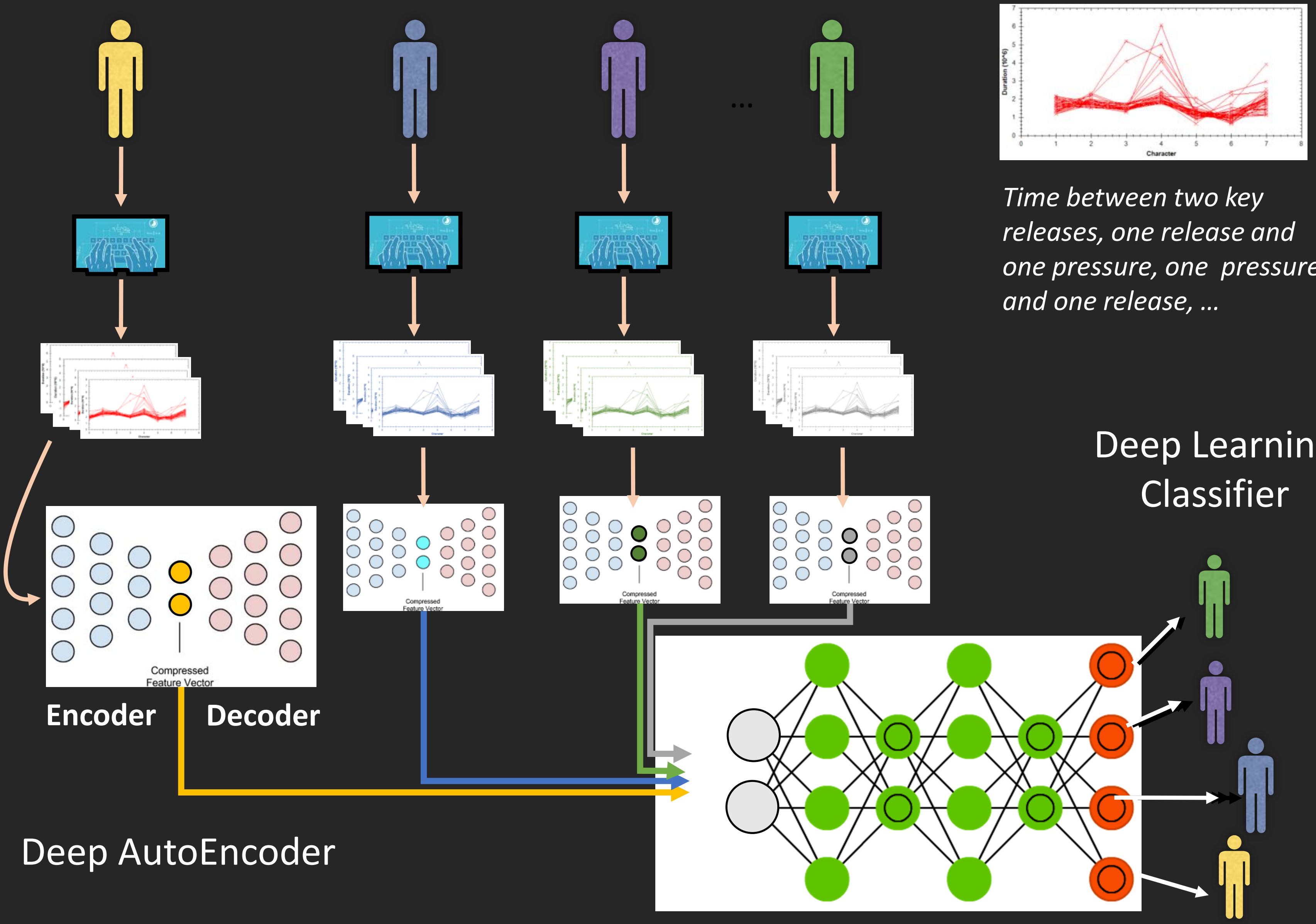
Identifying an individual based on her way of typing on a physical or virtual keyboard

eit Digital

Partners: FBK, IMDEA, REPLY COMMUNICATION VALLEY

Individual DeepKS Learner

C. Furlanello – MPBA Nov2017


...
Duration (10^-6)
Character
0
1
2
3
4
5
6
7
8
Compressed Feature Vector
Compressed Feature Vector
Compressed Feature Vector
Compressed Feature Vector

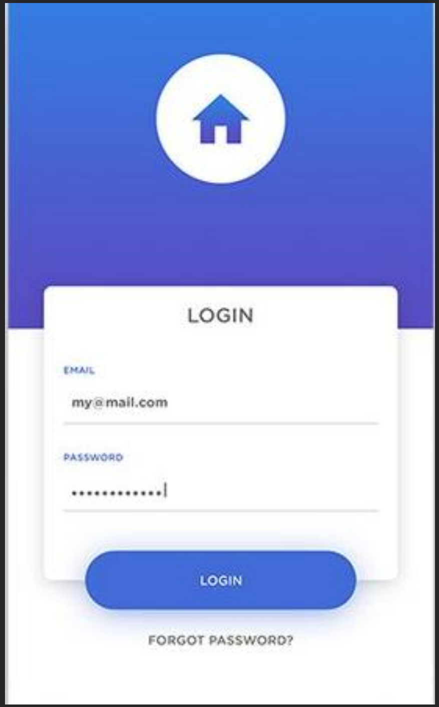
Time between two key releases, one release and one pressure, one pressure and one release, …

Deep Learning Classifier

Encoder **Decoder**

Deep AutoEncoder

Deep Learning as a Service



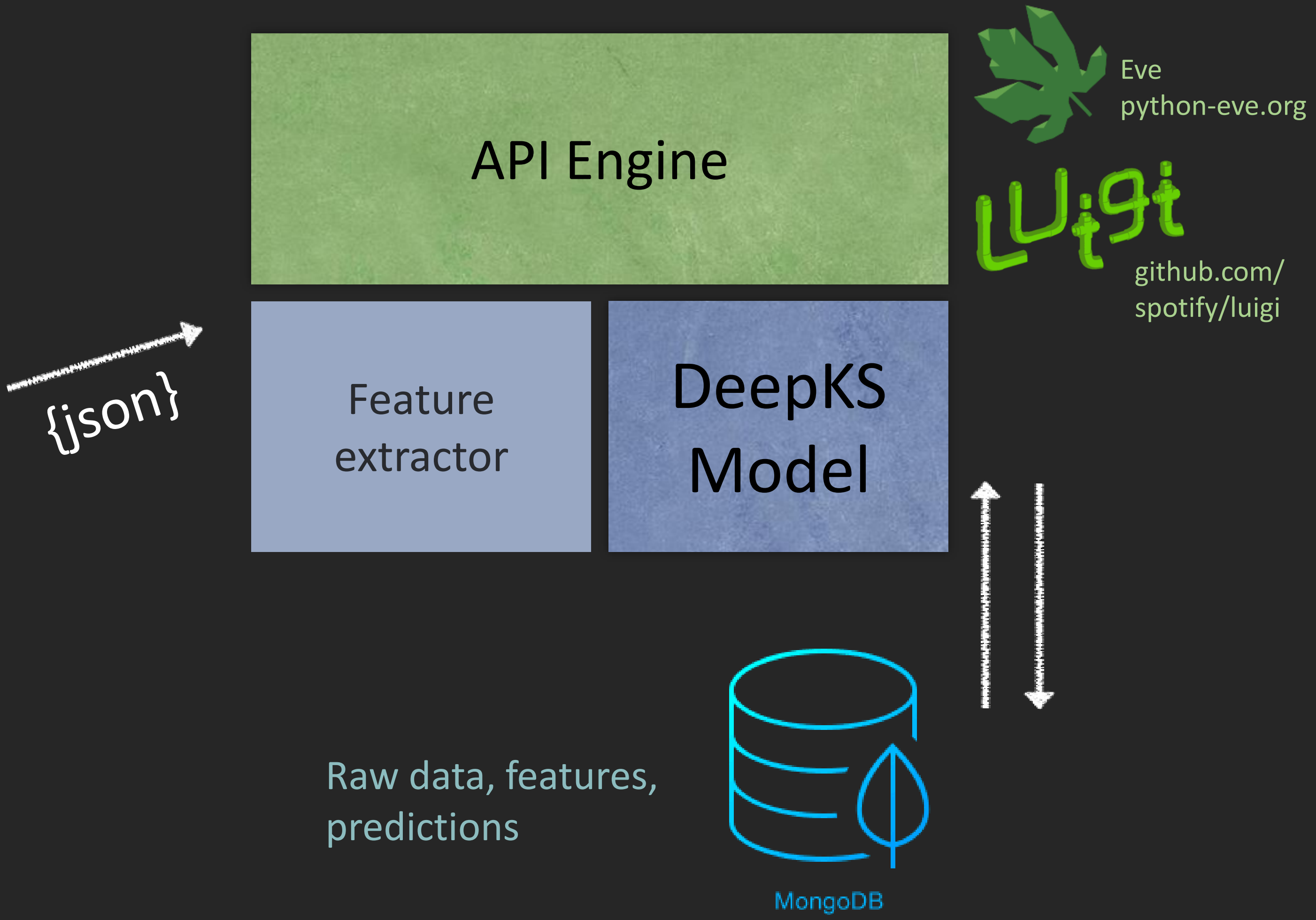
LOGIN

EMAIL

my@mail.com

PASSWORD

•••••••••••••



{json}

API Engine

Feature extractor

DeepKS Model

Eve
python-eve.org

luigi
github.com/spotify/luigi

Raw data, features, predictions

MongoDB

eit Digital

Partners: FBK, IMDEA, REPLY COMMUNICATION VALLEY

Execution

R&D competition driven by a new vision

Innovation in DL

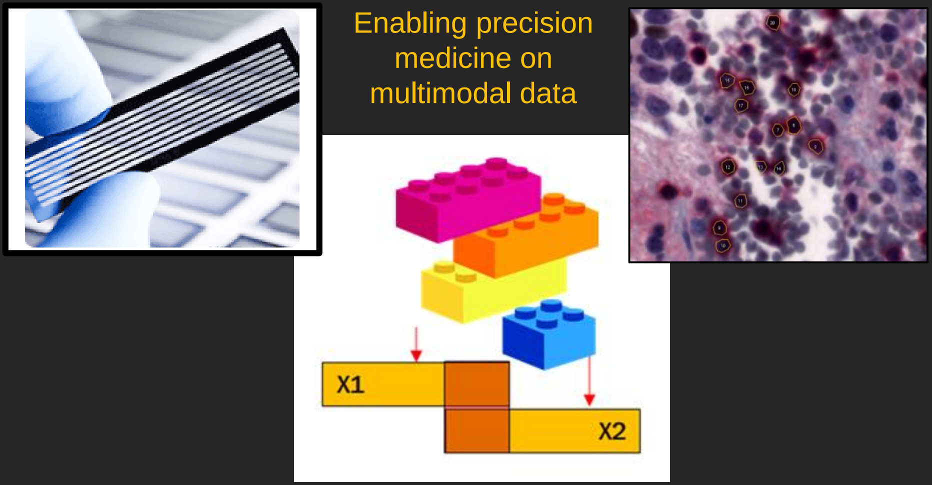
"run as a venture" within collaborations

Commit to grow a new generation of interdisciplinary researchers with strong entrepreneurship

Access to GPU resources, datasets, & "open science setups"

Resources for disruptive innovation

Unexpected Synergies



Enabling precision medicine on multimodal data

- **IOT for Smart Digital Industry**
- **Precision Medicine** for biotech/healthcare research
- **Environment:** big data and AI for life and food

Acceleration of Deep Learning technology on Agritech

Bayer and Monsanto – 2017

WebValley is the FBK summer school for data science and interdisciplinary research: close to 350 students from around the world (17-19y old) have attended the WebValley camps since its first edition in 2001.

In 2016 and 2017, the team developed a new Deep Learning solution for fruit quality control based on portable spectrometry and low cost images

Agritech as an accelerator of Precision Medicine: Deep Learning, cloud infrastructure (MS Azure), local GPU boxes, blockchain

START FAST, START EARLY

Thanks to G. Jurman, M. Chierici, C. Dolci, M. Cristoforetti, V. Maggior, Ernesto Arbitrio, G. Franch

C. Furlanello – MPBA Nov2017