SOSC 2018 Second International PhD School Open Science Cloud

Predictive models with Machine and Deep Learning:
a scientific view

D. Bonacorsi (University of Bologna)

SOSC'18 - Perugia, 17 Settembre 2018

Preface

Science needs conceptual representations of real phenomena

- → modelling

Operational value of a model relies in its predictive capabilities

- knowledge + data from past + math/stat → forecast future

Every scientific model is - at best - just a decent approximation of reality

- → model improvement and refinement (e.g. **seek and adopt new techniques**)

This is where Machine Learning (ML) and Deep Learning (DL)
knock on the doors of Science(s)

Outline

Predictive models are crucial in natural sciences and formal sciences

- Medical Science (e.g. predicting a disease, drug discovery, ..)
- Chemistry (e.g. predicting chemical reactions, ..)
- Bioinformatics (e.g. predicting protein structures, mining omic data, ..)
- Geosciences (e.g. predicting a rare catastrophic event, ..)
- Physics (e.g. predicting - and enforcing - a discovery of a new particle)
- Many more would deserve a discussion (Astronomy, Astrophysics, Earth Sciences, Climate, ..)

This talk will not review the theory of building predictive models, and/or ML/DL frameworks and algorithms (done well elsewhere in this School).

My goal today is to provide examples that highlight some level of similarity across science challenges, and how specific ML/DL tools might massively help a very diverse set of scientific disciplines.

Medical Science

Credits: Luca Antiga (Orobix CEO) for inspiration and (rearranged) material from his contribution at a AI for Industry event, Bologna, April 2017

Medical image analysis

ANNALS OF MEDICINE APRIL 3, 2017 ISSUE

A.I. VERSUS M.D.

What happens when diagnosis is automated?

By Siddhartha Mukherjee

THE NEW YORKER

Geoffrey Hinton, a computer scientist at the University of Toronto, speaks less gently about the role that learning machines will play in clinical medicine. Hinton—the great-great-grandson of George Boole, whose Boolean algebra is a keystone of digital computing—has sometimes been called the father of deep learning; it's a topic he's worked on since the mid-nineteen-seventies, and many of his students have become principal architects of the field today.

"I think that if you work as a radiologist you are like Wile E. Coyote in the cartoon," Hinton told me. "You're already over the edge of the cliff, but you haven't yet looked down. There's no ground underneath." Deep-learning systems for breast and heart imaging have already been developed commercially. "It's just completely obvious that in five years deep learning is going to do better than radiologists," he went on. "It *might* be ten years. I said this at a hospital. It did not go down too well."

Hinton's actual words, in that hospital talk, were blunt: "They should stop training radiologists now." When I brought up the challenge to Angela Lignelli-Dipple, she pointed out that diagnostic radiologists aren't merely engaged in yes-no classification. They're not just locating the embolism that brought on a stroke. They're noticing the small bleed elsewhere that might make it disastrous to use a clot-busting drug; they're picking up on an unexpected, maybe still asymptomatic tumor.

"Pretty good. The ending was a bit predictable."

Hinton now qualifies the provocation. "The role of radiologists will evolve from doing perceptual things that could probably be done by a highly trained pigeon to doing far more cognitive things," he told me. His prognosis for the future of automated medicine is based on a simple principle: "Take any old classification problem where you have a lot of data, and it's going to be solved by deep learning. There's going to be *thousands* of applications of deep learning." He wants to use learning algorithms to read X-rays, CT scans, and MRIs of every variety—and that's just what he considers the near-term prospects. In the future, he said, "learning algorithms will make pathological diagnoses." They might read Pap smears, listen to heart sounds, or predict relapses in psychiatric patients.

[H1]

G. Hinton, interviewed by The New Yorker

"I think that if you work as a radiologist you are like Wile E. Coyote in the cartoon (..) You're already over the edge of the cliff, but you haven't yet looked down. There's no ground underneath."

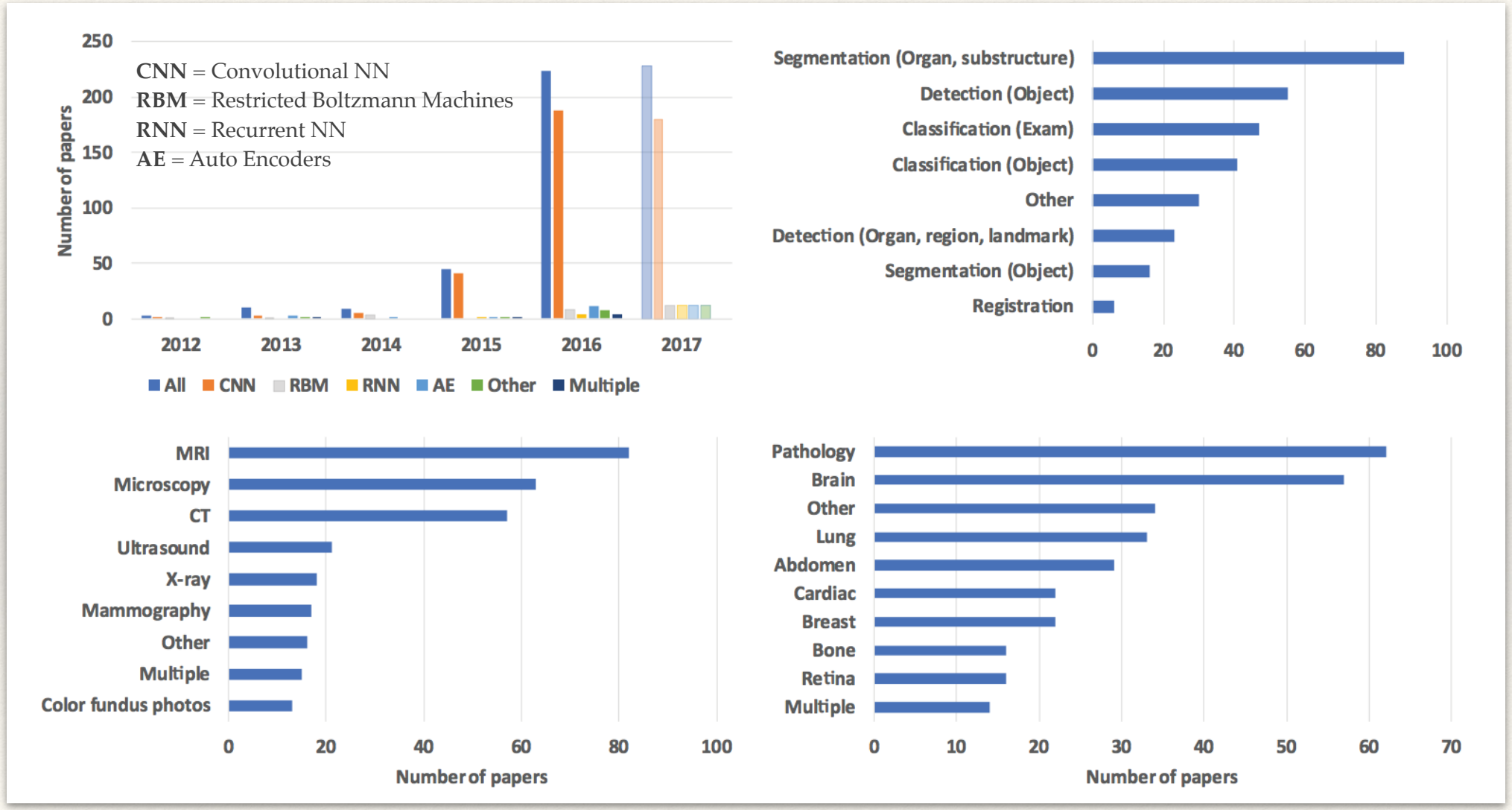
"It's just **completely obvious that in five years deep learning is going to do better than radiologists** (..) It might be ten years. I said this at a hospital. It did not go down too well."

"They should stop training radiologists now (..) The role of radiologists will evolve from doing perceptual things that could probably be done by a highly trained pigeon to doing far more cognitive things."

DISCLAIMER: "views are not my own"..

Deep Learning roars

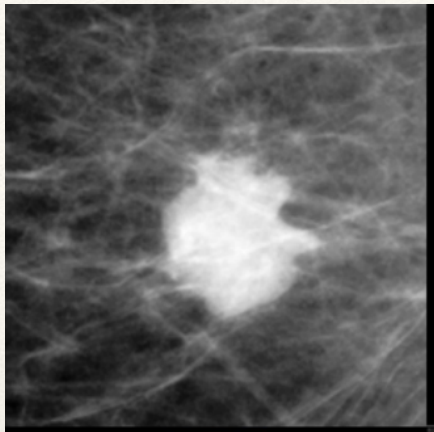
G. Litjens et al, "**A Survey on Deep Learning in Medical Image Analysis**", Jun 2017 [H2]

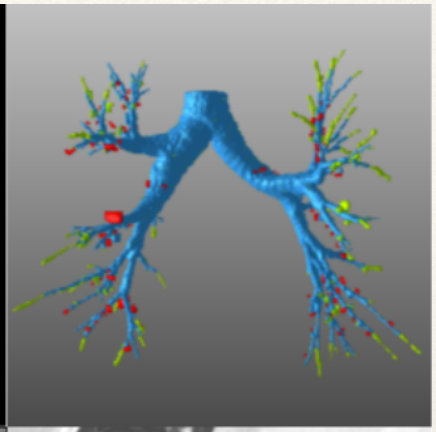


Breakdown of scientific papers by publication yr, task addressed, imaging modality, and application area

Vast set of application areas

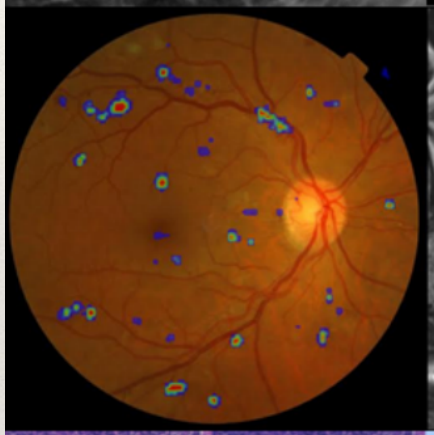
Mammography

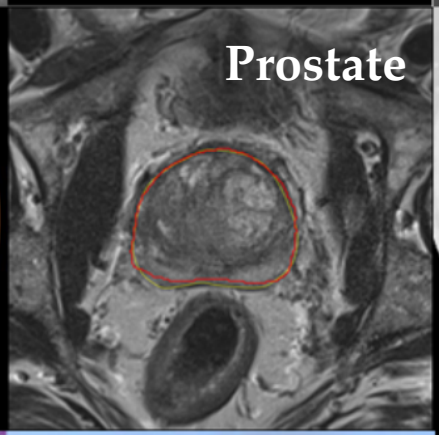


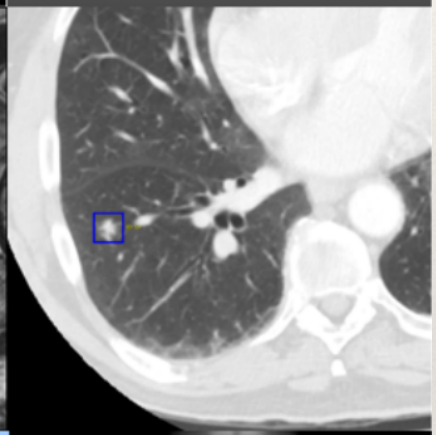


Airways

Diabetic retinopathy

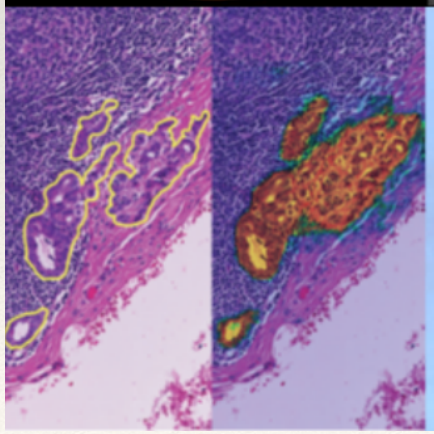


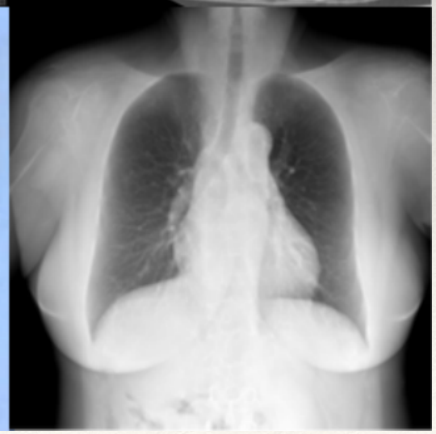




Lung nodule classification

Breast cancer metastases detection in lymph nodes



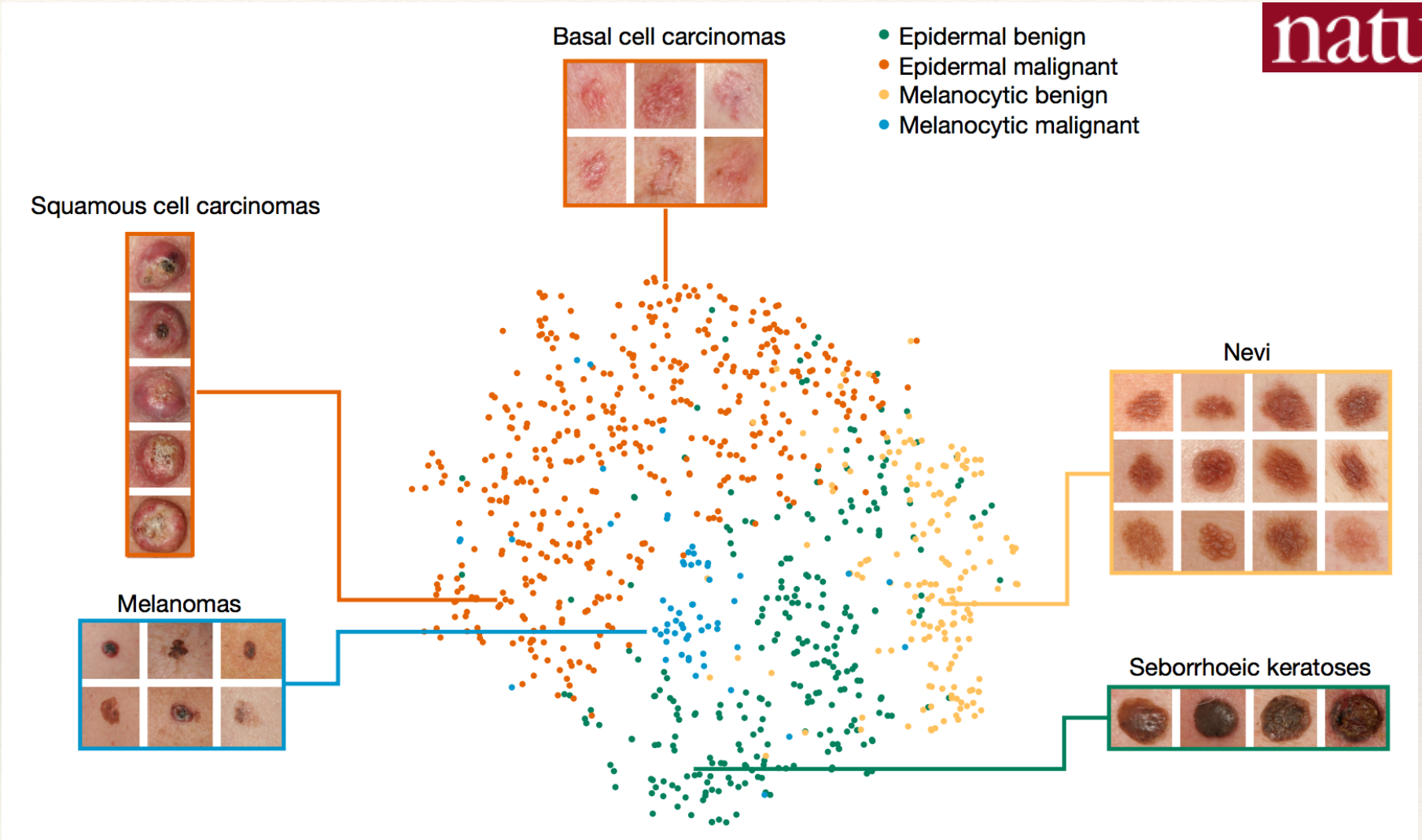


Bone suppression

[H2]

"Equivalence" demonstration in skin lesions [1/2]

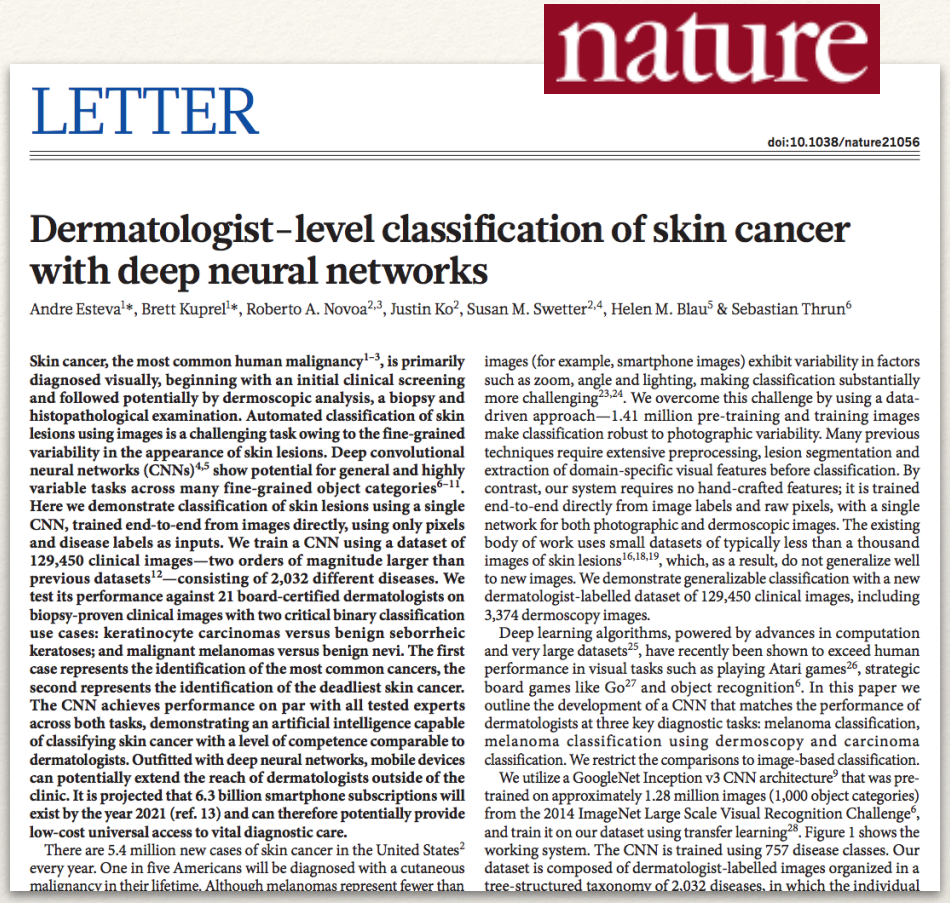
Demonstrated the **equivalence between a DL-based system and a pool of experts** in dermatology image classification



A. Esteva et al, "**Dermatologist-level classification of skin cancer with deep NN**", Feb 2017 [H3]

"Equivalence" demonstration in skin lesions [2/2]

"(…) Here we demonstrate classification of skin lesions using **a single CNN**, trained end-to-end from images directly, using only pixels and disease labels as inputs. We train a CNN using a dataset of **129,450 clinical images** - two orders of magnitude larger than previous datasets - consisting of 2,032 different diseases. We test its performance against **21 board-certified dermatologists on biopsy-proven clinical images** with two critical binary classification use cases: keratinocyte carcinomas versus benign seborrheic keratoses; and malignant melanomas versus benign nevi. The first case represents the *identification of the most common cancers*, the second represents the *identification of the deadliest skin cancer*. The CNN achieves **performance on par with all tested experts across both tasks, demonstrating an artificial intelligence capable of classifying skin cancer with a level of competence comparable to dermatologists.** *Outfitted with deep neural networks, mobile devices can potentially extend the reach of dermatologists outside of the clinic. It is projected that 6.3 billion smartphone subscriptions will exist by the year 2021 and can therefore potentially provide low-cost universal access to vital diagnostic care (…)"*



nature

LETTER

doi:10.1038/nature21056

Dermatologist-level classification of skin cancer with deep neural networks

Andre Esteva[1]*, Brett Kuprel[1]*, Roberto A. Novoa[2,3], Justin Ko[2], Susan M. Swetter[2,4], Helen M. Blau[5] & Sebastian Thrun[6]

Skin cancer, the most common human malignancy[1–3], is primarily diagnosed visually, beginning with an initial clinical screening and followed potentially by dermoscopic analysis, a biopsy and histopathological examination. Automated classification of skin lesions using images is a challenging task owing to the fine-grained variability in the appearance of skin lesions. Deep convolutional neural networks (CNNs)[4,5] show potential for general and highly variable tasks across many fine-grained object categories[6–11]. Here we demonstrate classification of skin lesions using a single CNN, trained end-to-end from images directly, using only pixels and disease labels as inputs. We train a CNN using a dataset of 129,450 clinical images—two orders of magnitude larger than previous datasets[12]—consisting of 2,032 different diseases. We test its performance against 21 board-certified dermatologists on biopsy-proven clinical images with two critical binary classification use cases: keratinocyte carcinomas versus benign seborrheic keratoses; and malignant melanomas versus benign nevi. The first case represents the identification of the most common cancers, the second represents the identification of the deadliest skin cancer. The CNN achieves performance on par with all tested experts across both tasks, demonstrating an artificial intelligence capable of classifying skin cancer with a level of competence comparable to dermatologists. Outfitted with deep neural networks, mobile devices can potentially extend the reach of dermatologists outside of the clinic. It is projected that 6.3 billion smartphone subscriptions will exist by the year 2021 (ref. 13) and can therefore potentially provide low-cost universal access to vital diagnostic care.

There are 5.4 million new cases of skin cancer in the United States[2] every year. One in five Americans will be diagnosed with a cutaneous malignancy in their lifetime. Although melanomas represent fewer than images (for example, smartphone images) exhibit variability in factors such as zoom, angle and lighting, making classification substantially more challenging[23,24]. We overcome this challenge by using a data-driven approach—1.41 million pre-training and training images make classification robust to photographic variability. Many previous techniques require extensive preprocessing, lesion segmentation and extraction of domain-specific visual features before classification. By contrast, our system requires no hand-crafted features; it is trained end-to-end directly from image labels and raw pixels, with a single network for both photographic and dermoscopic images. The existing body of work uses small datasets of typically less than a thousand images of skin lesions[16,18,19], which, as a result, do not generalize well to new images. We demonstrate generalizable classification with a new dermatologist-labelled dataset of 129,450 clinical images, including 3,374 dermoscopy images.

Deep learning algorithms, powered by advances in computation and very large datasets[25], have recently been shown to exceed human performance in visual tasks such as playing Atari games[26], strategic board games like Go[27] and object recognition[6]. In this paper we outline the development of a CNN that matches the performance of dermatologists at three key diagnostic tasks: melanoma classification, melanoma classification using dermoscopy and carcinoma classification. We restrict the comparisons to image-based classification.

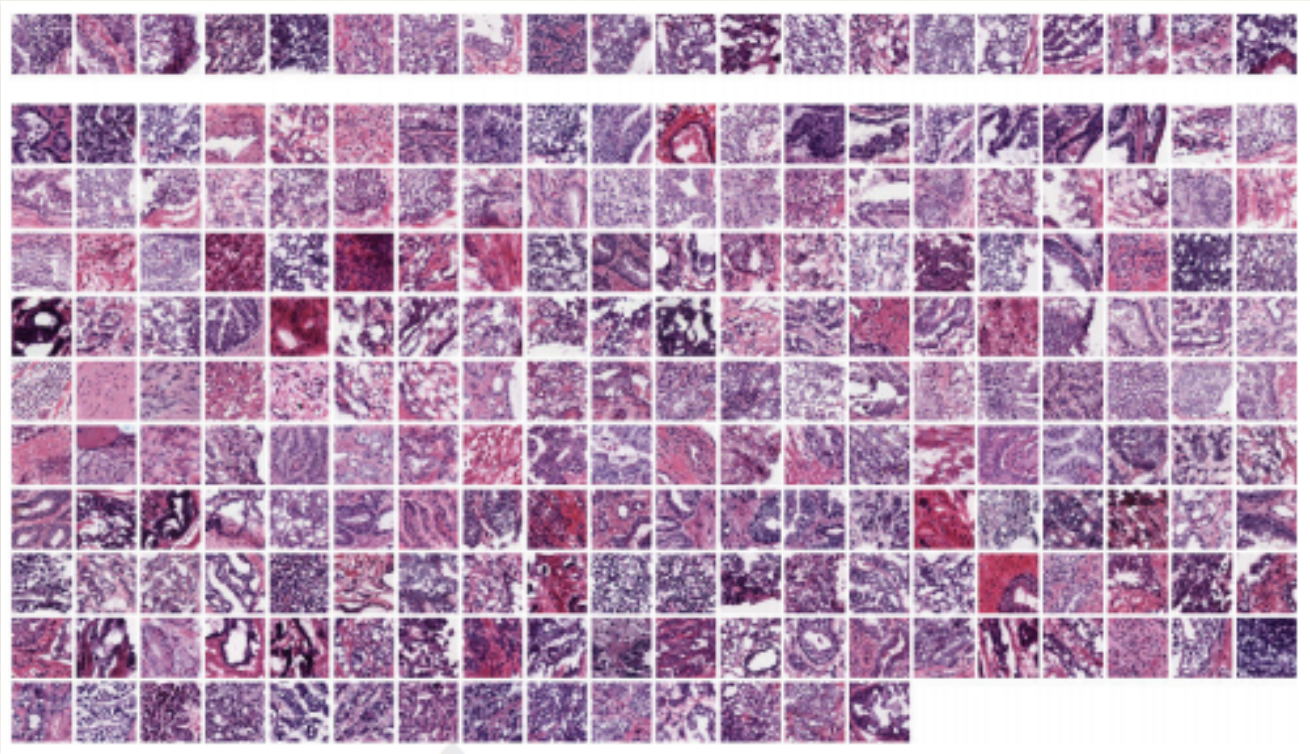
We utilize a GoogleNet Inception v3 CNN architecture[9] that was pre-trained on approximately 1.28 million images (1,000 object categories) from the 2014 ImageNet Large Scale Visual Recognition Challenge[6], and train it on our dataset using transfer learning[28]. Figure 1 shows the working system. The CNN is trained using 757 disease classes. Our dataset is composed of dermatologist-labelled images organized in a tree-structured taxonomy of 2,032 diseases, in which the individual

[H3]

Interesting "equivalence" demo, but - still - humans are able to do this.

Genetic mutation probability in prostate cancer [1/2]

"This is the first pipeline **predicting gene mutation probability in cancer from digitised H&E-stained microscopy slides**. To **predict whether or not the speckle-type POZ protein [SPOP] gene is mutated in prostate cancer**, the pipeline (i) identifies diagnostically salient slide regions, (ii) identifies the salient region having the dominant tumor, and (iii) trains **ensembles of binary classifiers that together predict a confidence interval of mutation probability.** *Through deep learning on small datasets, this enables automated histologic diagnoses based on probabilities of underlying molecular aberrations and finds histologically similar patients by learned genetic-histologic relationships*"



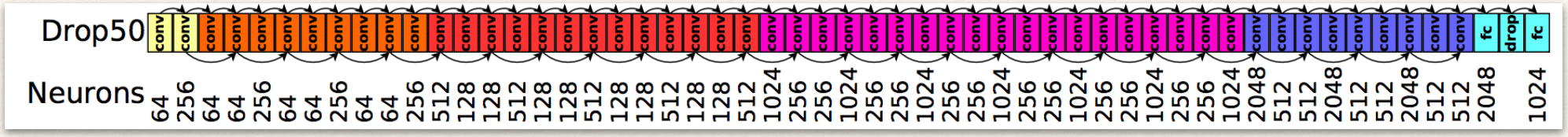
20 SPOP mutants

157 SPOP non-mutants

Source: TCGA cohort of frozen section images

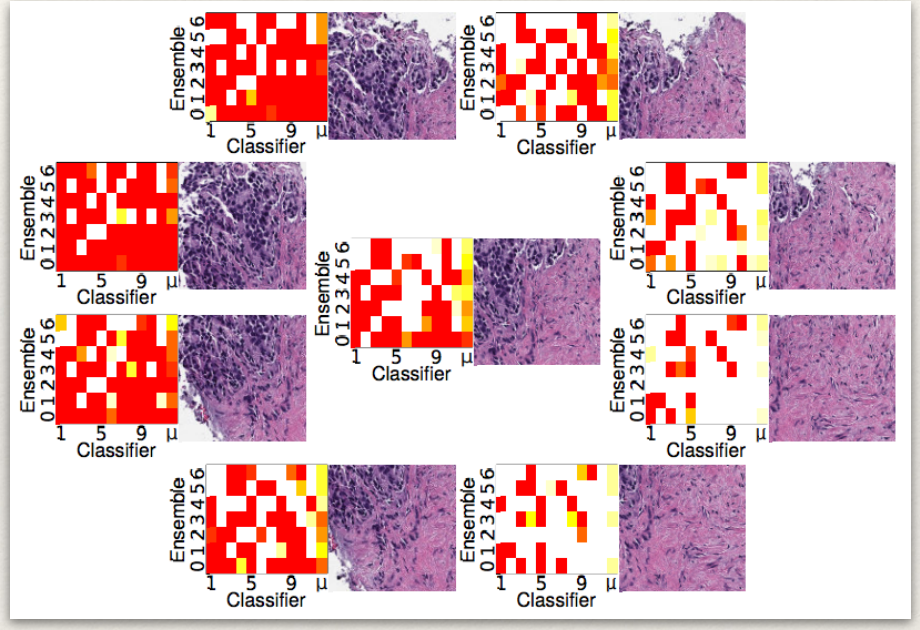
A.J. Schaumber et al, "**H&E-stained Whole Slide Image Deep Learning Predicts SPOP Mutation State in Prostate Cancer**". May 2017 [H4]

Genetic mutation probability in prostate cancer [2/2]



Customised ResNet-50 architecture, plus additional dropout and fully connected neuron layers (total 28,574 neurons)

RoI and patches with classification heatmaps. Strong SPOP mutation predictions is **red**, no such evidence is **white**. Weighted mean predictions are calculated per each patch, and combined. Meta-ensemble's SPOP mutation prediction here is 95% C.L.



This is interesting because **humans are NOT able to do this**

- experts might develop "intuitions" of this kind after decades of experience.. but this might become **an automated system always available in support to clinical activities**

Deep generative models

DL methods involving **discriminative models** are most commonly used (and successful) for classification tasks…

- based on back-propagation, dropout, piecewise linear units as activation functions… well-behaved GD

… increasing demand for **deep generative models**

- i.e. ways to use DL to directly generate a model that could be successfully applied to e.g. compression, denoising, inpainting and/or texture synthesis, semi-supervised learning, unsupervised feature learning, …

.. *with the latter being a much bigger challenge than the former*!

- initial generative models (e.g. restricted/deep Boltzmann machines, denoising autoencoders, ..) are probabilistic and based on a parametric specification of a probability distribution function. Training of such models requires the maximization of the log-likelihood, a function that is usually computationally intractable, with the additional complication of the activation functions..
- several alternative (deep) generative models have been suggested, which do not require the explicit representation of the likelihood while being able to generate samples from the desired distribution

Latest class of non-parametric approaches for deep generative models is known as Generative Adversarial Network (GAN)

- generative models are estimated via an adversarial process. More in [IG1]

I. Goodfellow et al. "**Generative Adversarial Nets**" [IG1]

Artificially-intelligent drug discovery engines [1/2]

www.impactjournals.com/oncotarget/ Oncotarget, 2017, Vol. 8, (No. 7), pp: 10883-10890

Research Paper

The cornucopia of meaningful leads: Applying deep adversarial autoencoders for new molecule development in oncology

Artur Kadurin[1,2,3,4], Alexander Aliper[2], Andrey Kazennov[2,7], Polina Mamoshina[2,5], Quentin Vanhaelen[2], Kuzma Khrabrov[1], Alex Zhavoronkov[2,6,7]

[1]Search Department, Mail.Ru Group Ltd., Moscow, Russia

[2]Pharmaceutical Artificial Intelligence Department, Insilico Medicine, Inc., Emerging Technology Centers, Johns Hopkins University at Eastern, Baltimore, Maryland, USA

[3]Big Data and Text Analysis Laboratory, Kazan Federal University, Kazan, Republic of Tatarstan, Russia

[4]St. Petersburg Department of V.A. Steklov Institute of Mathematics of the Russian Academy of Sciences, Petersburg, Russia

[5]Department of Computer Science, University of Oxford, Oxford, UK

[6]The Biogerontology Research Foundation, Trevissome Park, Truro TR4 8UN, UK

[7]Moscow Institute of Physics and Technology, Dolgoprudny, Russia

Correspondence to: *Alex Zhavoronkov,* ***email:*** *alex@insilicomedicine.com*

Keywords: *generative adversarian networks, adversarial autoencoder, deep learning, drug discovery, artificial intelligence*

Received: *June 14, 2016* ***Accepted:*** *November 24, 2016* ***Published:*** *December 22, 2016*

ABSTRACT

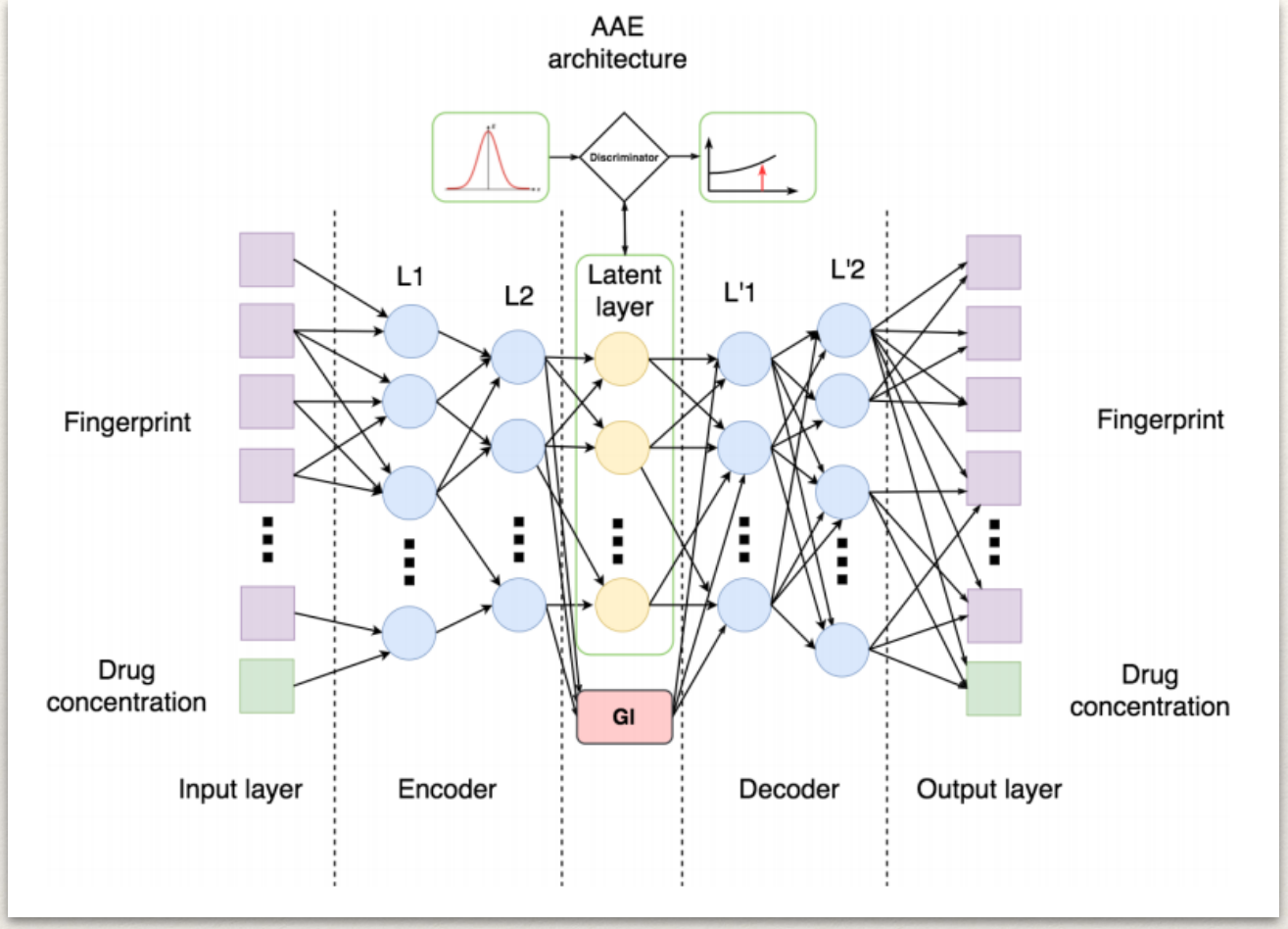
Recent advances in deep learning and specifically in generative adversarial networks have demonstrated surprising results in generating new images and videos upon request even using natural language as input. In this paper we present the first application of generative adversarial autoencoders (AAE) for generating novel molecular fingerprints with a defined set of parameters. We developed a 7-layer AAE architecture with the latent middle layer serving as a discriminator. As an input and output the AAE uses a vector of binary fingerprints and concentration of the molecule. In the latent layer we also introduced a neuron responsible for growth inhibition percentage, which when negative indicates the reduction in the number of tumor cells after the treatment. To train the AAE we used the NCI-60 cell line assay data for 6252 compounds profiled on MCF-7 cell line. The output of the AAE was used to screen 72 million compounds in PubChem and select candidate molecules with potential anti-cancer properties. This approach is a proof of concept of an artificially-intelligent drug discovery engine, where AAEs are used to generate new molecular fingerprints with the desired molecular properties.

"Recent advances in deep learning and specifically in GANs have demonstrated surprising results in generating new images and videos upon request even using natural language as input."

"In this paper we present **the first application of generative adversarial autoencoders (AAE) for generating novel molecular fingerprints with a defined set of parameters**. (…) This approach is a proof of concept of an artificially-intelligent drug discovery engine, where *AAEs are used to generate new molecular fingerprints with the desired molecular properties.*

A. Kadurin et al, "**The cornucopia of meaningful leads: Applying deep adversarial autoencoders for new molecule development in oncology**", Dec 2016 [H5]

Artificially-intelligent drug discovery engines [2/2]



Architecture of the 7-layer Adversarial Autoencoder (AAE) used in the aforementioned study.

Encoder consists of 2 consequent layers L1 and L2 with 128 and 64 neurons. In turn, **decoder** consists of 2 layers L'1 and L'2 comprising 64 and 128 neurons. Latent layer consists of 5 neurons, one of which is Growth Inhibition percentage (GI) and the other 4 are discriminated with normal distribution.

(..) "As an input and output the AAE uses a **vector of binary fingerprints and concentration of the molecule**. In the latent layer we also introduced a neuron responsible for growth inhibition percentage, which **when negative indicates the reduction in the number of tumor cells after the treatment**. To train the AAE we used the NCI-60 cell line assay data for 6252 compounds profiled on MCF-7 cell line. The output of the AAE was used to **screen 72 million compounds in PubChem and select candidate molecules with potential anticancer properties**."

One "non technical" thought (-> ethics)

Having an automated system has been a convenience so far

- and experts always dominated..

Now, not in the medical job at large, but in specific pattern matching tasks, these **systems easily beat humans**.

Ethical implications?

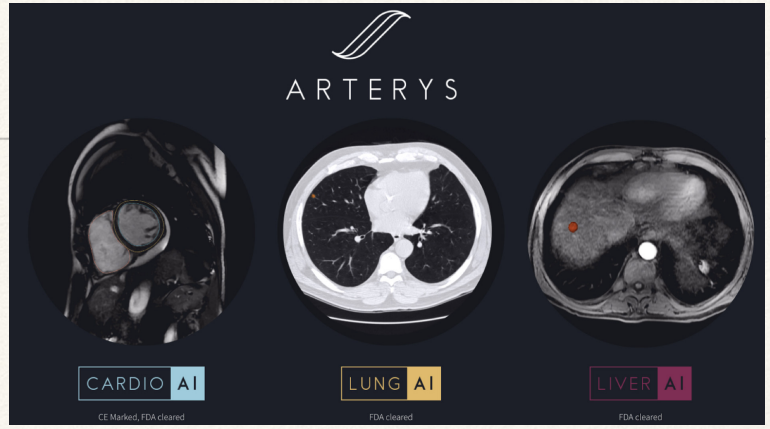
- "is it ethical?"
- or, even, "will it become non-ethical not to use an automated system?"

Other implications

- Many! E.g. future of jobs. E.g. how will this transform regulatory rules, FDA, …

(perhaps, to be seen *as a human-machine collaboration and not a competition..)*

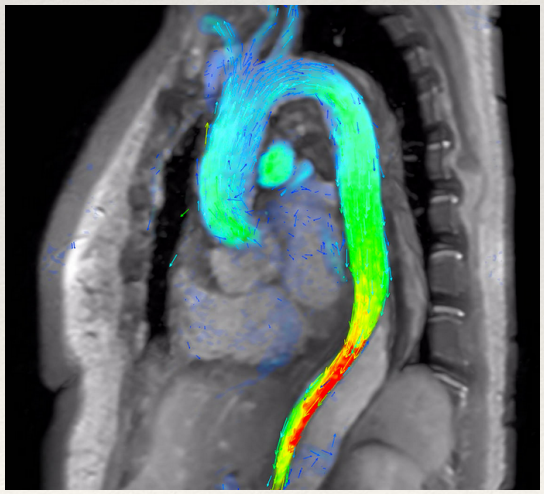
FDA-approved DL systems



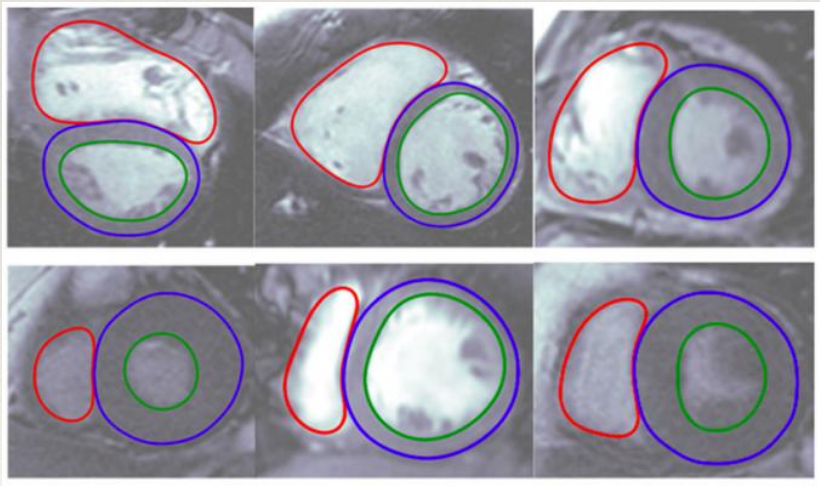
Example. *Artemys*, first start-up that had a DL-method (segmentation of cardiac valves, ..) approved by FDA

- quoted here only for one reason: perhaps a milestone indicative of just one company's momentum in applying AI to advance medical imaging accuracy

FDA 510(k) approval Nov'16
4D flow (velocity vectors + time) from MRI scans



FDA 510(k) approval Feb'17
Cardiac valves segmentation, validated on ~1k patients



[*] 510(k) = premarket submission made to FDA to demonstrate that the device to be marketed is at least as safe and effective - that is, substantially equivalent - to a legally marketed device

Outlook on role(s) of ML in Medical Science

ML prospers with Big Data, and Medical science is yielding large amount of heterogeneous data daily

- R&D, physicians and clinics, patients, caregivers, etc.

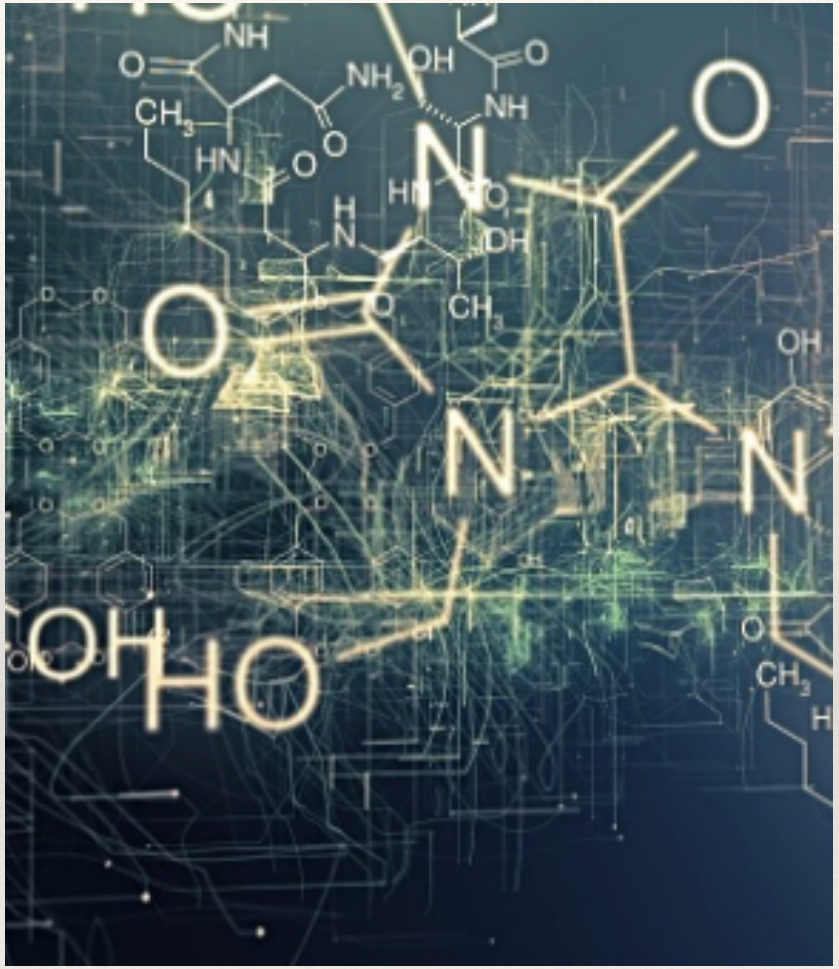
Gains:

- better decision-making, optimised innovation, improved efficiency of research/clinical trials, creation of new tools for physicians, consumers, ..

A (non exhaustive) list of applications of ML in Medical Science - for those interested:

- Disease Identification/Diagnosis → *a quick example in this talk*
- Personalised Treatment/Behavioural Modification
- Drug Discovery/Manufacturing → *a quick example in this talk*
- Clinical Trial Research
- Smart Electronic Health Records
- Epidemic Outbreak Prediction
- Radiology and Radiotherapy → *a quick example in this talk*
- …

Chemistry



Predicting chemical reactions

Predicting the course and major products of arbitrary reactions is a fundamental problem in **Chemistry**

- chemists address this in a variety of ways, from synthesis design to reaction discovery

Basically, two different approaches:

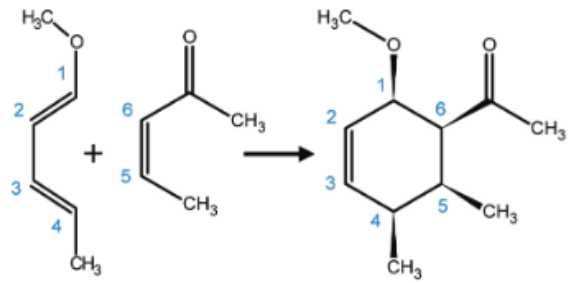
- Write a system of rules (so-called "rule-based experts systems")
  - limitation: very tedious, non-scalable, limited coverage
- Learn the rules from Big Data
  - traditional inductive ML may not suffice, as you lack sufficient data to be implemented

Chemistry towards ML [1/2]

E.g. “Reaction Explorer”, a system to predict organic chemical reactions based on a knowledge base of >1500 manually composed reaction transformation rules.

Product prediction for a Diels-Alder reaction using the accompanying “SMIRKS” transformation rule.

Huge expressive power of the rules to enforce regioselectivity, stereospecificity, stereoselectivity of the reaction (e.g. carbon 1 preferentially assumes an ortho position with respect to carbon 6, etc).



[C1]

[*:10]/[CX3;$(*[O,N,S]):1]=[C:2]\-[C:3]=[CX3:4]/[*:11].
[*:12]\[CX3:5]=[CX3:6]/[$(*=[O,N,S]),$(C#N):13]>>
[*:10][CX4;@:1]1-/[C:2]=[C:3]\-[CX4;@@:4]([*:11])[CX4;@:5]([*:12])[CX4;@@:6]1[*:13]

A new method uses ML [C2]:

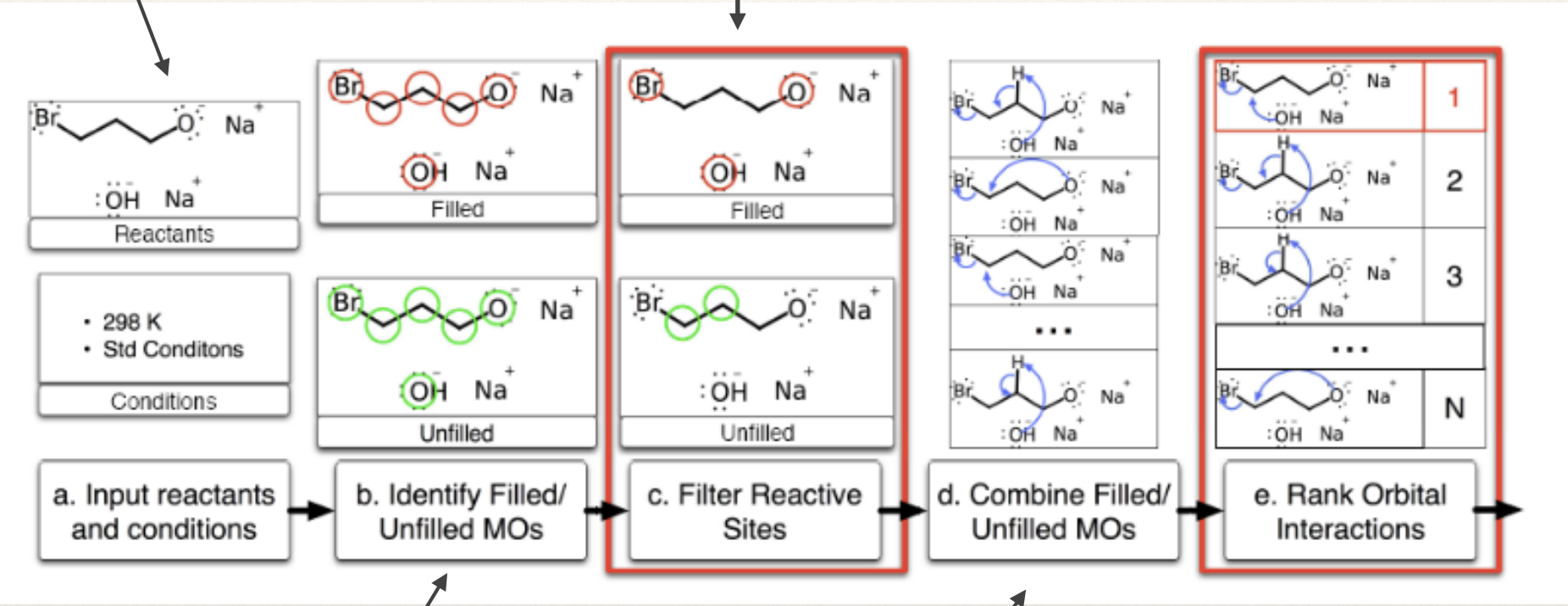
- describe single reactions as interactions between coarse approximations of molecular orbitals (MOs). Use topological and physicochemical attributes as descriptors.
- Use an existing rule-based system (Reaction Explorer) to derive a restricted chemistry dataset consisting of 1630 full multi-step reactions with 2358 distinct starting materials and intermediates, associated with 2989 productive mechanistic steps and 6.14 million unproductive mechanistic steps
- Use ML: formulate identifying productive mechanistic steps as a statistical ranking, information retrieval, problem: given a set of reactants and a description of conditions, learn a ranking model over potential filled-to-unfilled MO interactions such that the top ranked mechanistic steps yield the major products.

Chemistry towards ML [2/2]

A user inputs the reactants and conditions

Highly sensitive reactive site **classifiers** are trained and used to filter out the vast majority of unreactive sites, pruning the space of potential reactions.

[C2]



Potential e⁻ donors and acceptors are identified using coarse approximations of electron filled and electron unfilled MOs.

Reactions are enumerated by pairing filled and unfilled MOs.

A ranking model is trained and used to order the reactions, where the best ranking one or few represent the major products.

Bioinformatics

Bioinformatics

"Bioinformatics" = answer biological questions using tools from mathematics, statistics and *computer science*

- advanced computational tools → boost in collection and analysis of biological data

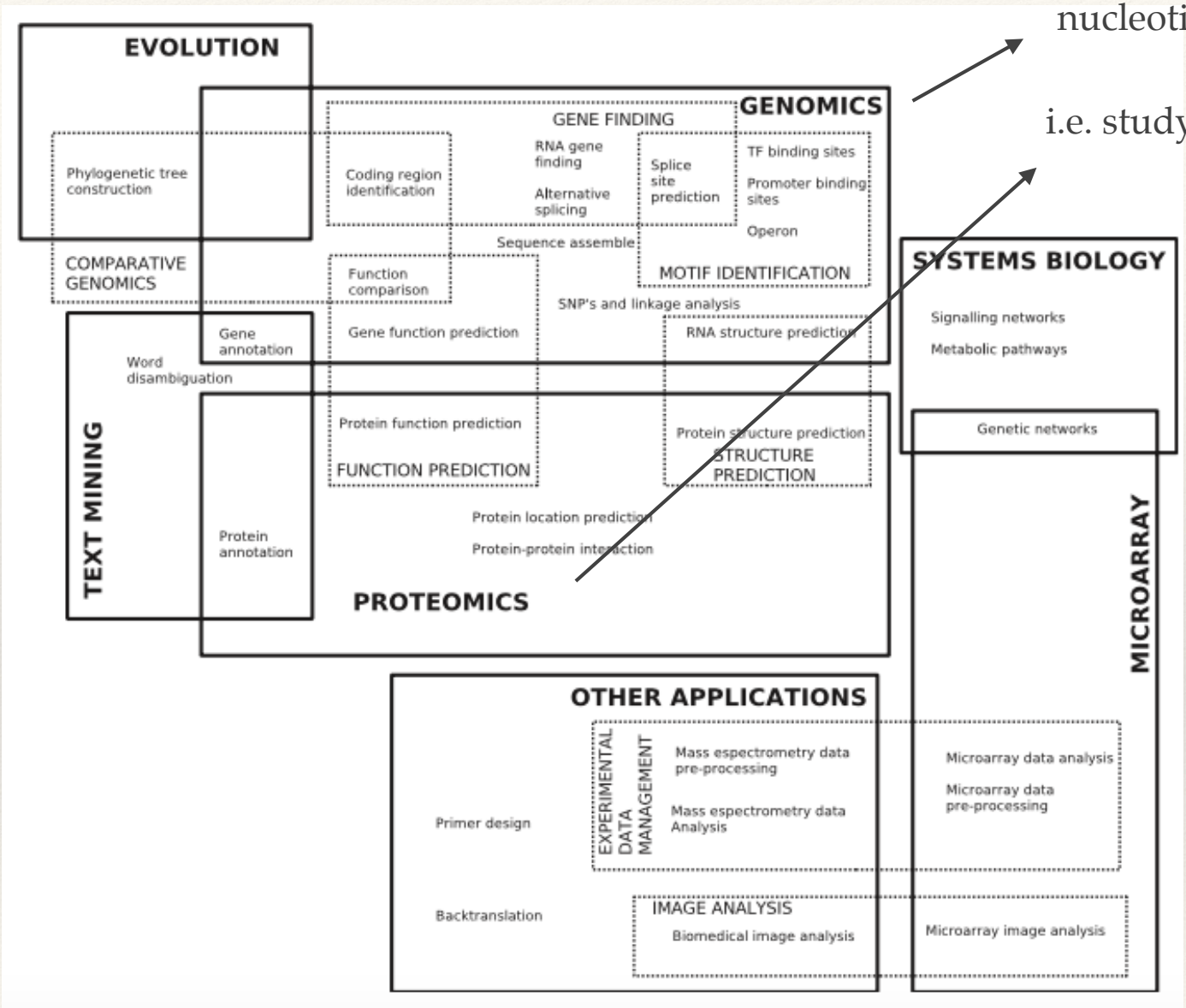
Biological "sequences" represent a large portion of biological data

- large size of the sequences + numerous possible features → **strong need of powerful analysis methods and tools**

ML in Bioinformatics

i.e. study of nucleotide chains

i.e. study of proteins



[B2]

Very complex mapping of ML methods to biological tasks..

ML in genomics and proteomics

Genomics

- one of the most important domains in bioinformatics, as genes contain all the information
  - From genome sequences, **location and structure of the genes** are extracted. The regulatory elements and non-coding RNA genes are identified. Sequence information is also used for **gene function and RNA secondary structure prediction**.
- **Big (Bio-)data**: # of sequences available is increasing exponentially. From 1982 to present, the # of bases in GenBank has ~doubled every ~18 months. Large data volume is richness for ML..

Proteomics

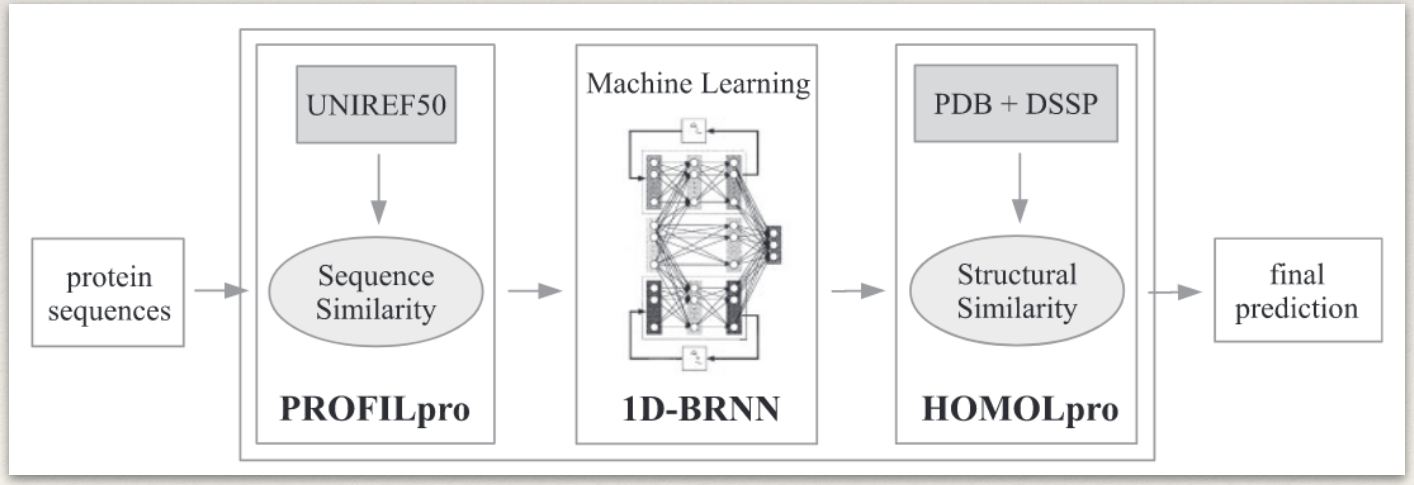
- proteins transform the information in the genes into life
- The goal is **protein structure prediction**: their 3D structure is a key feature in their functionality (evolution, structure and function). But proteins are very complex macromolecules with thousands of atoms and bounds. Hence, the number of possible structures is huge, and protein structure prediction is **a very complicated combinatorial problem where optimisation techniques are required**. This is where computational methods are needed.

In both genomics and proteomics, ML techniques are applied for protein function prediction

Example: mining omic data

Most effective **predictors of protein 3D structure** (able e.g. to capture secondary structures) nowadays are a combination of methods

- Not only sequence profiles at the input, but also sequence similarity and structural(e.g. similarity to sequences in the Protein Data Bank used to infer annotations at the output level), then you can use Bidirectional Recursive Neural Networks



General workflow of (selected) predictors [B1]. Sequence and structural similarity analyses are performed by stand-alone modules (those named "*pro") and BRNN models are trained to predict the features from the profiles and combined in an ensemble

ML on Microarrays and in Systems biology

Microarray:

- essays as the best known (despite not the only one) domain where bio-data is collected
- complex experimental data need to be pre-processed (i.e. modified to be suitably used by ML algos), then the data analysis method depends on what it is being looked for
- most typical ML applications are on **expression pattern identification, classification and genetic network induction**

Systems biology:

- very complex to model the life processes that take place inside the cell
- ML helpful ingredients in **modelling biological networks** (especially genetic networks), **signal transduction networks** and **metabolic pathways**

ML on Evolution and Text-mining

Evolution

- ML used especially in **phylogenetic tree reconstruction**
- traditionally: these schematic representations of organisms' evolution were constructed according "only" to different features (morphological features, metabolic features, etc.)
- today: great amount of genome sequences available → phylogenetic tree reconstruction algos based on the **comparison between different genomes**, made by means of multiple sequence alignment, where optimisation techniques are very useful

Text-mining

- data proliferation → text mining techniques useful for knowledge extraction and organisation, and are becoming popular as a side effect of (big) biodata
- applied in **functional annotation**, **cellular location prediction** and **protein interaction analysis** - more in [B3]

Which ML approach(es) for bioinformatics?

You find in literature application of: (not exhaustive list)

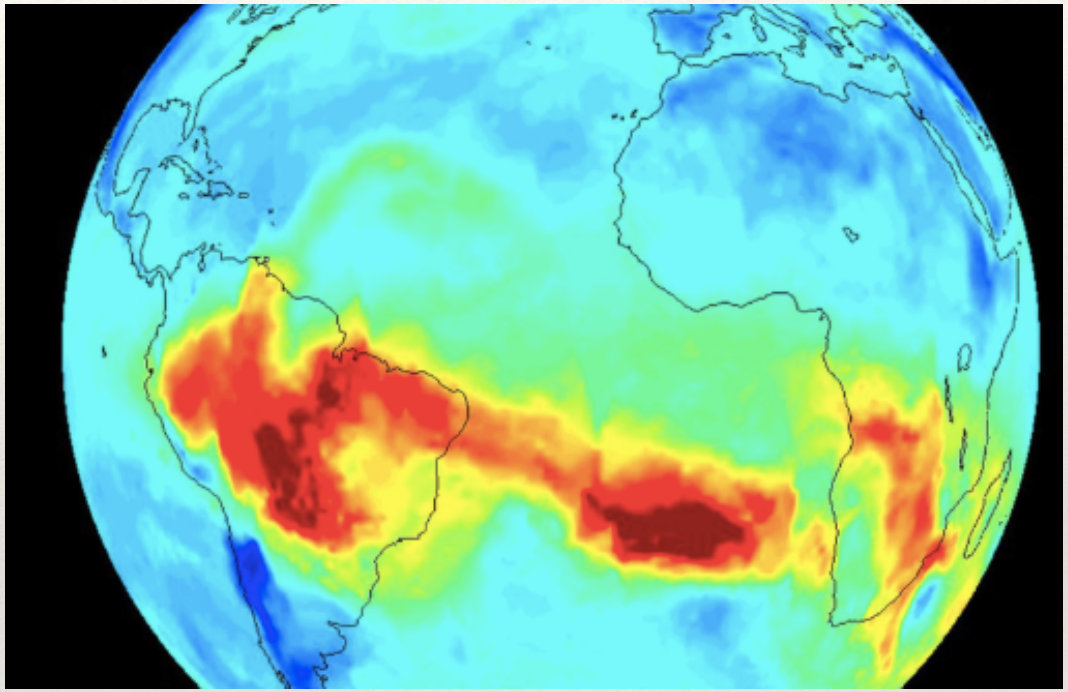
Bayesian classifiers, logistic regression, discriminant analysis, classification trees, nearest neighbour, neural networks, support vector machines, ensembles of classifiers, partitional clustering, hierarchical clustering, mixture models, hidden Markov models, Bayesian networks and Gaussian networks, ..

Few *examples*:

- e.g. identification of specific biological sequence segments with **NN**, **Bayesian classifiers**, **decision trees**, and **SVM**
- when standard ML approaches fail, focus goes to **feature generation**, **feature selection**
- also, **clustering algos** are used to group structurally related biological sequences

Bioinformatics and ML/DL is a very active and interesting field!

Geoscience



Geoscience(s)

Earth's major interacting components are complex dynamic systems

- e.g. litho- sphere, biosphere, hydrosphere, and atmosphere

Their states perpetually keep changing in space and time, creating a balance of mass and energy

- e.g., layers in oceans, ions in air, minerals and grains in rock, land covers on the ground

All interact with each other through **complex and dynamic geoscience processes**

- e.g. rain falling on Earth's surface and nourishing the biomass; sediments depositing on river banks and changing river course; magma erupting on sea floor and forming islands..

Geo-data comes mainly from 2 broad categories of sources:

- 1. observational data collected via sensors (space, sea, land)
- 2. simulation data from physics-based models of the Earth system.

Big (geo-)Data

Geosciences are a field of great societal relevance, requiring solutions to urgent problems that humanity is facing

- impact of climate change; air pollution; increased risks to infrastructures by disasters (such as hurricanes); modelling future availability and consumption of water, food, and mineral resources; identifying factors responsible for earthquake, landslide, flood, and volcanic eruption

Research is **extremely complex**, as it is at the confluence of various disciplines

- e.g. physics, geology, hydrology, chemistry, biology, ecology..

The **Big Data** era impacted geosciences too, which **became a data-rich field**

- better sensing technologies (e.g., remote sensing satellites and deep sea drilling vessels)
- improvements in computational resources for running large-scale simulations of Earth system models
- Internet-based democratisation of data, enabling collection, storage, processing of data on crowd-sourced and distributed environments such as cloud platforms

Several unique challenges that are seldom found in other sciences, mostly related to the typical sources of geoscience data and their properties. In this scenario, ML offer immense potential to contribute to problems in Geosciences

Hard to use ML on geo-data

Several characteristics of geo-data and geoscience applications limit the usefulness of traditional ML algos for knowledge discovery, e.g.:

the nature of geoscience processes

- *objects with amorphous boundaries* (e.g. waves, flows, ..). E.g. advanced fluid segmentation and fluid feature characterisation are needed
- *space-time structure*. Land cover labels (e.g. forest, desert, urban, ..) require high resolution in space, and can change over time. High correlations. Cannot use ML methods that assume independent and identically distributed variables.
- *high dimensionality*. Earth system incredibly complex, huge # of potentially correlated variables (e.g. detection of land cover changes requires analysis of multiple remote sensing variables)
- *rare processes*. Most catastrophic events would be the most useful to predict. But historical occurrences are few, hence issues with the skew (imbalance) between the rare and not-rare classes.

geoscience data collection

- *multi-resolution data*. E.g. sources like satellite sensors or in-situ measurements are associated to varying spatial and temporal resolutions, sampling rate, accuracy, uncertainty. Need algos that can identify patterns at multiple resolution.
- *noise, incompleteness, and uncertainty in Data*

scarcity of samples and ground truth.

- *small sample size*. Issue from both reliable sensor-based data (e.g. satellites only since the 1970s), and rarity of some major events (landslides, tsunamis, forest fires, M>6 earthquakes). A killer for ML/DL approaches.
- *paucity of labeled samples with gold-standard ground truth*. High-quality measurements need very expensive apparatuses (e.g. low-flying airplanes), or expensive and time-consuming field operations. In addition, some geoscience processes (e.g. subsurface flow of water) do not have ground truth at all (so complex that exact state of the system is never fully known). Underfitting vs overfitting issues in ML.

Possible ML directions for Geosciences [1/2]

1. Characterising Objects and Events

- characterise and identify objects (e.g. weather fronts, atmospheric rivers); analyse patterns in geo-data objects to study events (e.g. tornado-genesis)
- beyond using hand-coded features (size, shape), ML can help in automated detection from data with improved performance using pattern mining techniques, provided that can account for the s+t properties of geo-data
  - done e.g. for spatio-temporal patterns in sea surface height data [G2], resulting in the creation of a global catalogue of ocean eddies

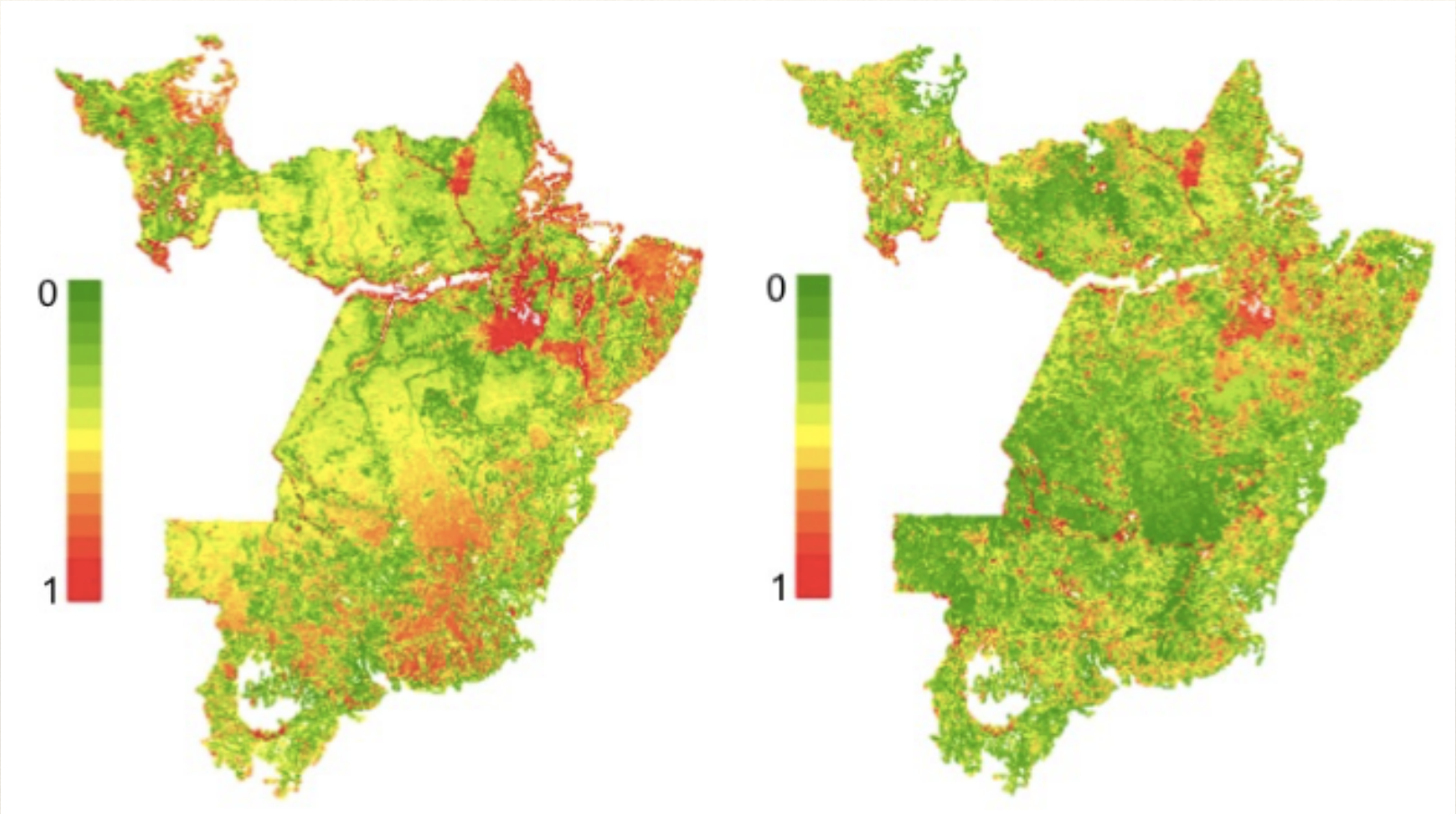
2. Estimating Geoscience Variables from Observations

- supervised ML can help to infer critical geoscience variables that are difficult to monitor directly (e.g. use data about other variables collected via satellites and ground-based sensors, or simulations). E.g. multi-task learning (MTL) is used (which improve generalisation by leveraging the domain-specific information contained in the training signals of related tasks)

- to address the non-stationary nature of some geo-data (e.g. climate), online machine learning is used - dynamically adapting to new patterns in the data - to predict e.g. temperatures. This approach outperforms the traditional, non-adaptive (multi-model) mean over expert predictors

Baseline ML

Multitask ML



Performance improvement in estimating forest cover in 4 states of Brazil ("green" is "better performance) [G3]

Possible ML directions for Geosciences [2/2]

3. Long-term Forecasting of Geoscience Variables

- traditionally, run physics-based model simulations that encode geo-processes using state-based dynamical systems (current state determined by previous plus observations). Now: attacking as a time-series ML regression problem (e.g. hidden Markov models, ..)
- even more complex are long-term forecasts for rare events (due to few data, sparsity, etc). Promising is transfer learning, as model training on a present task (with sufficient # of training samples) can be used to improve prediction performance on a future task (with limited # of training samples)

4. Mining Relationships in Geoscience Data

- find relationship among different geo-physical processes. One class of such relationship in the climate domain is the "teleconnections" (pairs of distant regions highly correlated in climate variables such as sea level P or T)
- huge potential of data-driven approaches here, that can sift through vast volumes of observational and model-based geoscience data and discover interesting patterns

5. Causal Discovery and Causal Attribution

- discover cause-effect relationships. Traditionally, causality tools are used, e.g. bivariate Granger analysis or multi-variate Granger analysis using vector autoregression (VAR) models (the latter, together with Pearl's framework, not yet so common though)
- reinforcement learning and other stochastic dynamic programming approaches that can solve decision problems with ambiguous risk are promising directions that geoscientists are pursuing

Raise of DL in Geosciences

DL ability to automatically extract relevant features from the data

- huge potential in geoscience (difficult to otherwise build hand-coded features for objects/events/relationships)

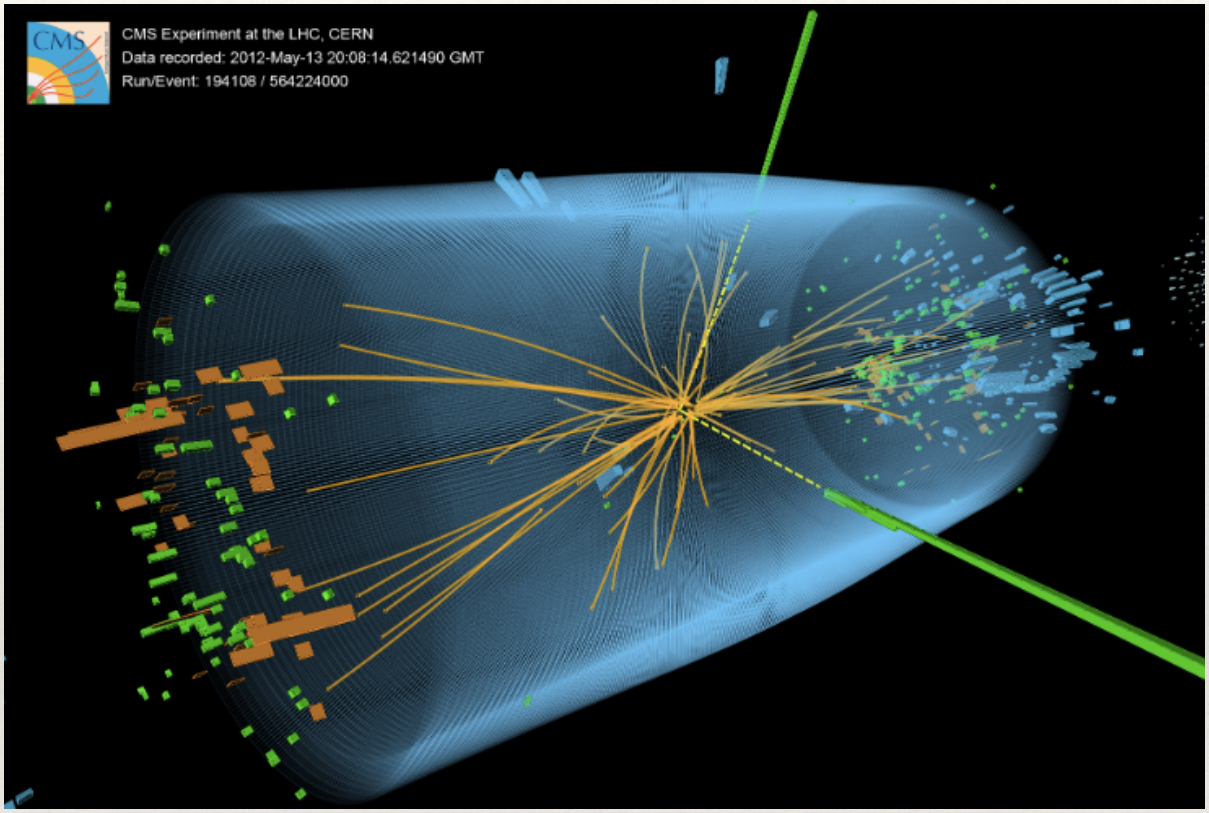
The space-time nature of geo-data raises some similarity with problems like computer vision and speech recognition, where DL excels

- → frameworks such CNNs and RNNs are used more and more
  - CNN already used for detecting extreme weather events from climate model simulations *[G4]*
  - RNN-based frameworks (such as LSTM models) have been explored for mapping plantations in Southeast Asia from remote sensing data *[G5]*
- DL systems explored also for downscaling outputs of Earth system models and generating climate change projections at local scales *[G6]*, and for classifying objects in high-resolution satellite images

Warning: availability of large volumes of labeled data has been a key factor behind the DL success. Paucity of labeled samples in geosciences is hence an issues, limiting the effectiveness of traditional DL methods.

- need to develop novel DL frameworks for geoscience (e.g. using domain-specific information of physical processes?)

Physics



DISCLAIMER: focus mostly on High Energy Physics (HEP) with particle accelerators.

High Energy Physics (HEP)

HEP focus is the study of **fundamental interactions** among **elementary particles**

- quarks and leptons as building blocks
- aiming at a complete understanding of *microcosm* and *macrocosm*

HEP physicists **create matter**

- they need to observe and study it beyond the ordinary one, hence they create matter in the states it existed fractions of seconds after the Big Bang

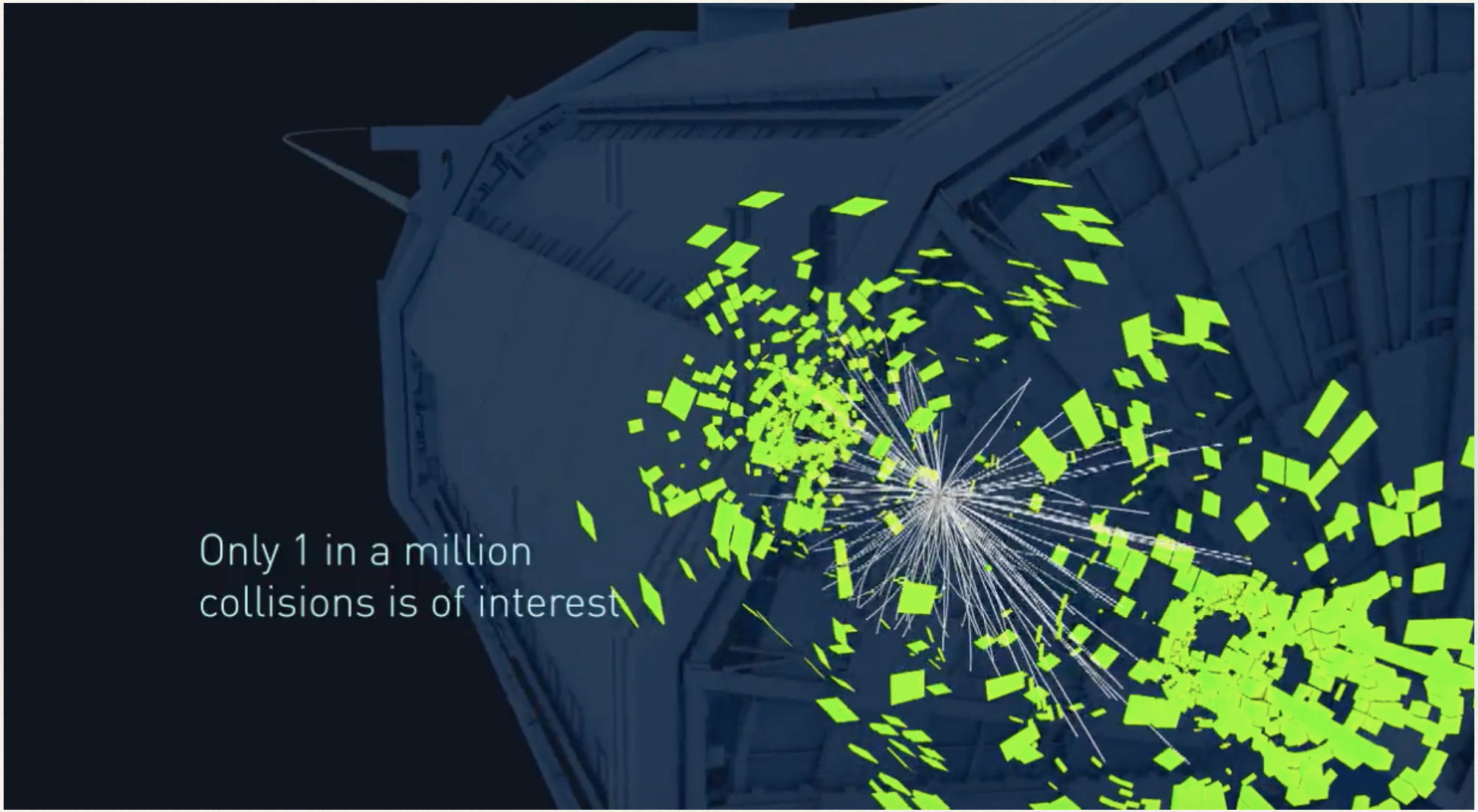
HEP physicists' instruments are **particle accelerators** plus large and complicated **particle detectors** around interaction points

- build and operate accelerators, accelerate particles to collisions, measure fragments that fly through the active volumes of the detectors → physics!
- or - in the case of Astrophysics - the Universe is a "natural particle accelerator"

This is amazingly fascinating and beautiful. And so complicated..

HEP with LHC at CERN

Check out full video at: https://videos.cern.ch/record/1541893



Innovation is hard(er in Big Science)

HEP community is **at the frontier of computing technologies**

- (apart from the obvious WWW born at CERN..) HEP has driven Grid Computing worldwide

But HEP community is extremely **large**, work on **long timescales**, and some **inertia** in a otherwise flexible adoption of new paradigms can be observed

- Current generation experimental programmes last *decades*. Long planning, long construction time, long operation by huge collaborations (~1000s of scientists)

Software and Computing experts from previous generation of experiments **pioneered studies employing ML** and laid the ground for the emergence of ML as an essential tool for HEP

But HEP timescales are decades, while ML/DL evolution timescale is years (or less..).

Today, important focus is in cross-discipline fertilisation (cultural and technical)

- Incorporating the "latest greates" new ML/DL tools in experiments that are finally taking data after decades of construction and large investments.. while maintaining the scientific rigour required in particle physics analyses.. in such a huge scientific environment.. all this presents some unique (not only technical!) challenges and opportunities

ML for HEP

Very wide field of **supervised ML** (mostly), e.g. training algorithms to classify data as signal or background by studying existing labeled (possibly Monte Carlo) data.

There are some HEP groups contributing to ML research worldwide, but most ML usage in HEP - as in most other sciences - is not research on ML

- HEP community is building domain-specific applications on top of existing toolkits and ML algorithms developed by computer scientists, data scientists, and scientific software developers from outside the HEP world

Work is also being done to understand where HEP problems do not map well onto existing ML paradigms and how these problems can be recast into abstract formulations of more general interest

ML algorithms in HEP

BDTs/ANNs typically used to classify particles and events

- they are also used for regression, e.g. to obtain the best estimate of particle's energy based on the measurements from several detectors

ANNs being used for a while in HEP, then.. → rise of **DNNs**

- particularly promising when there is a large amount of data and features, as well as symmetries and complex non-linear dependencies between inputs and outputs

Different types of NNs used in HEP:

- fully-connected (**FCN**), convolutional (**CNN**), recurrent (**RNN**) network
- additionally, NNs are used in the context of Generative Models, when a NN is trained to mimic multidimensional distributions to generate any number of new instances. Variational AutoEncoders (**VAEs**) and more recent Generative Adversarial Networks (**GANs**) are two examples of such generative models used in HEP.

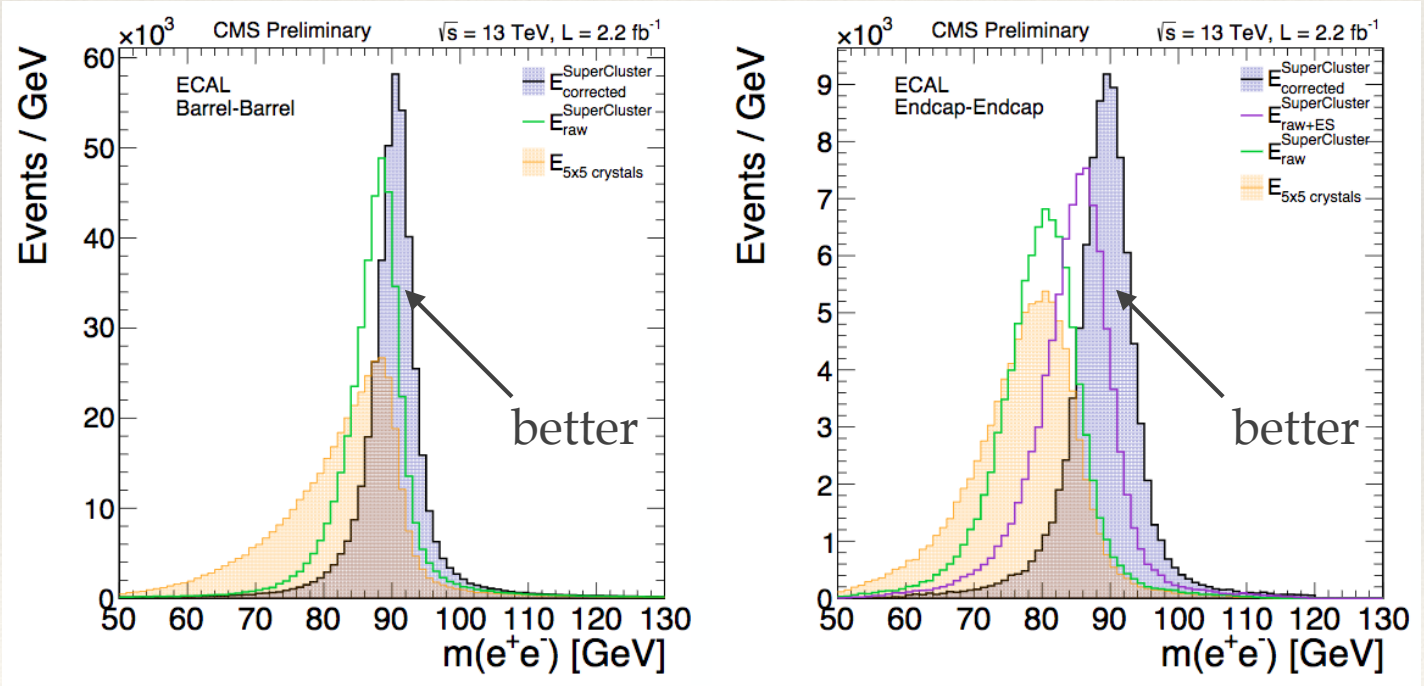
Plus, ML algorithms devoted to time-series analysis

- in general not relevant for HEP where events are independent from each other
- however, growing interest in these algorithms for HEP-related sequential non-collision data, e.g. for **Data Quality** and **Computing Infrastructure monitoring** (as well as those physics processes and event reconstruction tasks where time is an important dimension)

Particle properties: energy resolution

Using ML to improve the determination of particle properties is now commonplace in **all LHC experiments**

- E.g. energy deposited in calorimeters is recorded by many sensors, which are clustered to reconstruct the original particle energy. **CMS** is training **BDTs** to learn corrections using all information available in the various calorimeter sensors - thus resulting in a sizeable improvement in resolution



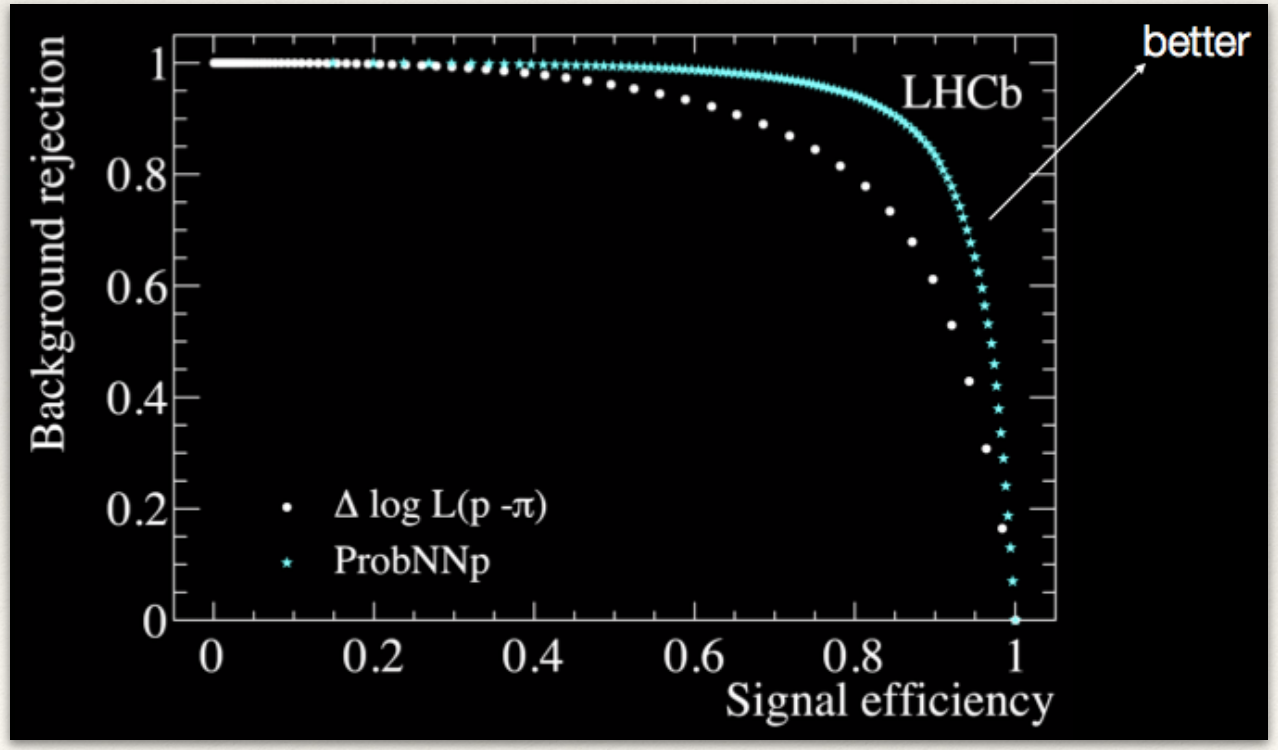
Improvements to the Z→e+e- energy scale and resolution from the incorporation of more sophisticated clustering and cluster correction algorithms (energy sum over the seed 5x5 crystal matrix, bremsstrahlung recovery using supercluster, inclusion of pre-shower energy, **energy correction using a multivariate algorithm**)

[2015 ECAL detector performance plots, CMS-DP-2015-057. Copyright CERN, reused with permission]

Particle ID

Similarly, ML is commonly used to identify particle types

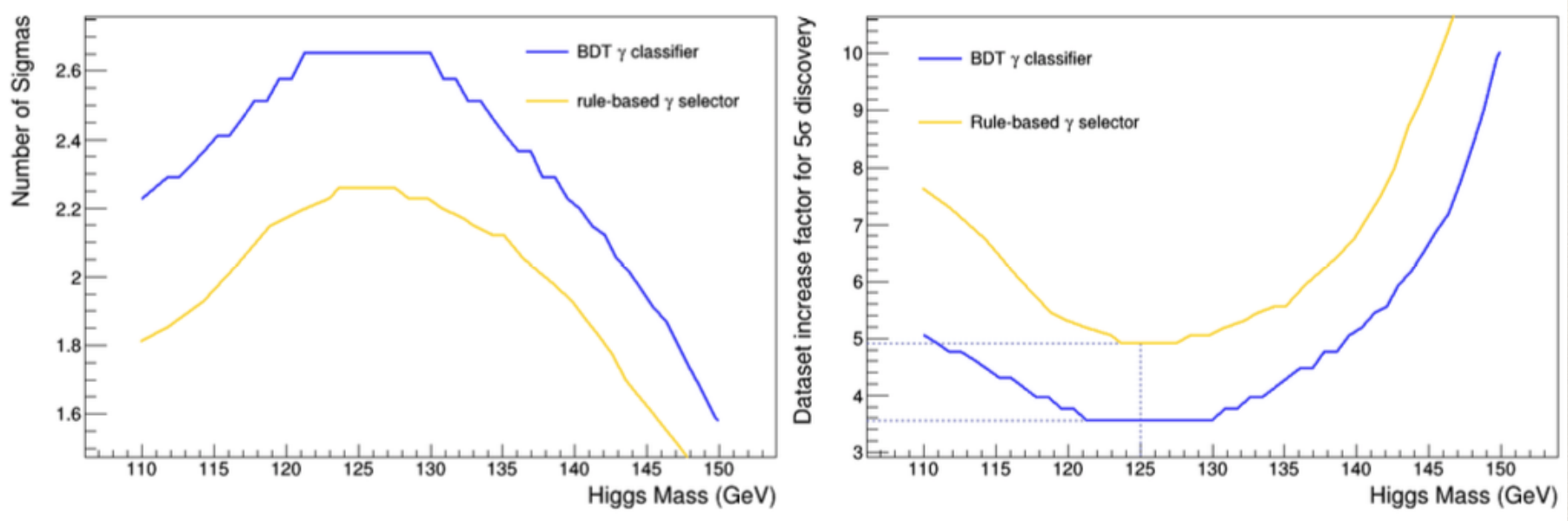
- e.g. **LHCb** uses **NN**s trained on O(30) features from all its subsystems, each of which is trained to identify a specific particle type
- ~3x less mis-ID bkg /particle. Estimates indicate that more advanced algorithms may reduce bkg by another ~50%



[courtesy: M.Williams]

Discovery of the Higgs boson

ML played a key role in the discovery of the Higgs boson, especially in the diphoton analysis by **CMS** where ML (used to improve the resolution and to select/categorise events) increased the sensitivity by roughly the equivalent of collecting ~50% more data.



[courtesy M.Pierini]

Higgs was not supposed to be discovered as early as in 2012

- Given how machine progressed, a possible discovery was expected by end 2015 / mid 2016

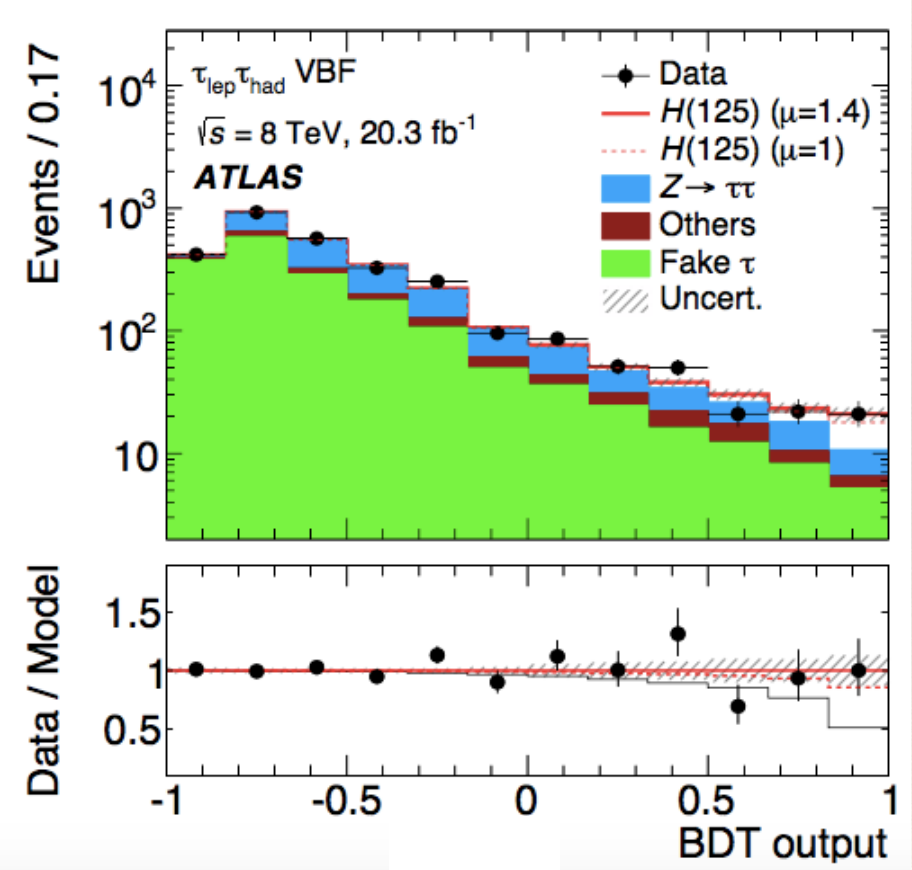
This was possible (and a Nobel was awarded) in advance, thanks (also) to ML

Study of Higgs properties

E.g. analysis of τ leptons at LHC complicated as they decay before being detected + loss of subsequently produced neutrinos + bkg from Z decays

- e.g. **ATLAS** divided the data sample into 6 distinct kinematic regions, and in each a **BDT** was trained using 12 weakly discriminating features → improved sensitivity by ~40% vs a non-ML approach

Also part of the 2014 Higgs ML **Kaggle** challenge

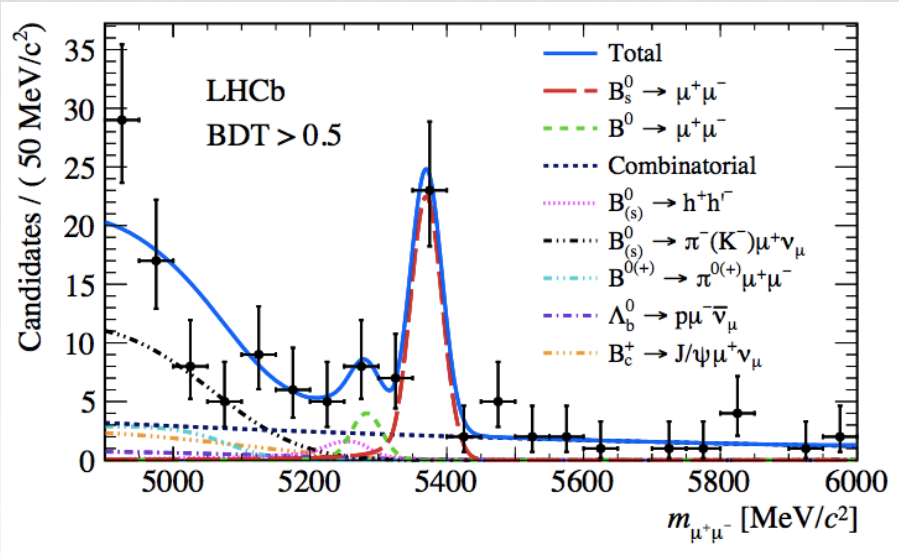


[arXiv:1501.04943]

High-precision tests of the SM

CMS and **LHCb** were the first to find evidence for the $B^0_s \to \mu^+\mu^-$ decay with a combined analysis (as rare as ~ 1 / 300 billion pp collisions..)

- **BDT**s used to reduce the dimensionality of the feature space - excluding the mass - to 1 dimension, then an analysis was performed of the mass spectra across bins of BDT response
- decay rate observed is consistent with SM prediction with a precision of ~25%, placing stringent constraints on many proposed extensions to the SM
- To obtain the same sensitivity without ML, LHCb as a single experiment would have required ~4x more data. Just one of many examples of high-precision tests of the SM at the LHC where ML can dramatically increase the power of the measurement



Mass distribution of the selected $B^0 \to \mu^+\mu^-$ candidates with BDT > 0.5.

[arXiv: 1703.05747]

Trigger

Crucial trade-off in algorithm complexity and performance under strict inference time constraints

E.g. **ATLAS/CMS** each only keep **about 1 in every 100 000 events**, and yet their data samples are each still **about 20 PB/yr**

- ML algorithms have already been used very successfully for rapid event characterisation
- adoption depth vary across experiments, but the increasing event complexity at High Luminosity LHC will require more sophisticated ML solutions and its expansion to more trigger levels

A critical part of this work will be to understand which ML techniques allow us to maximally exploit future computing architectures

Trigger (cont'd)

E.g. **CMS** employs ML in its trigger hardware to better estimate the momentum of muons

- inputs to the algorithm are discretised to permit encoding the ML response in a large look-up table that is programmed into FPGAs

E.g. **LHCb**, many of the reactions of greatest interest do not produce striking signatures in the detector, making it necessary to thoroughly analyse high-dimensional feature spaces in real time to efficiently classify events

- LHCb used a **BDT** for 2 years, then a MatrixNet algorithm
- ML now almost ubiquitous in LHCb Trigger. 70% of all persisted data is classified by ML algorithms. All charged-particle tracks are vetted by **NN**s.
- LHCb estimated that reaching the same sensitivity as a recent LHCb search for the dark matter on 2016 data, would have required collecting data for 10 yrs without the use of ML

Computing resource optimisations

Industrial-scale data samples collected by e.g. LHC experiments produce **non-collisions metadata from which actionable insights can be extracted**

- results of logging while running LHC Run-1/2 operations of complex Grid systems

ML techniques have begun to play a crucial role in increasing the efficiency of computing resource usage for LHC experiments since few years

- e.g. predicting which data will be accessed the most to a-priori optimise data storage at Grid computing centres via pre-placement, or perform WAN path optimisation based on user access historical patterns (done/in-progress primarily, but not only, in **LHCb** and **CMS**)
- e.g. monitoring data transfer latencies over complex network topologies, using ML to identify problematic nodes and predict likely congestions (in progress by **CMS**)

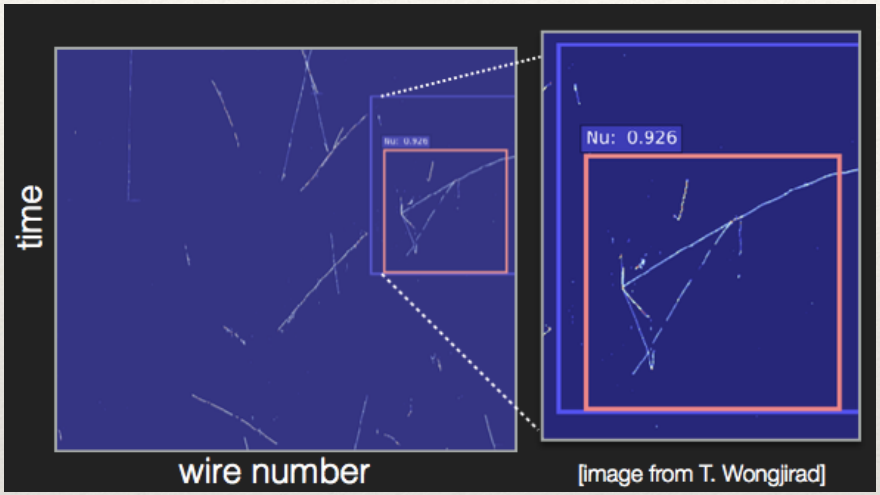
Current approach is that ML informs the choices of the computing operations teams

- this might be the basis of fully adaptive models in the next future

CNNs for neutrinos

MicroBooNE has managed to train **CNN**s that can locate neutrino interactions within an event in their LArTPC, identify objects and assign pixels to them

- CNN perfect to identify objects in an image (translational invariant feature learning), and sensitive volumes are large due characteristics of neutrino interaction with matter



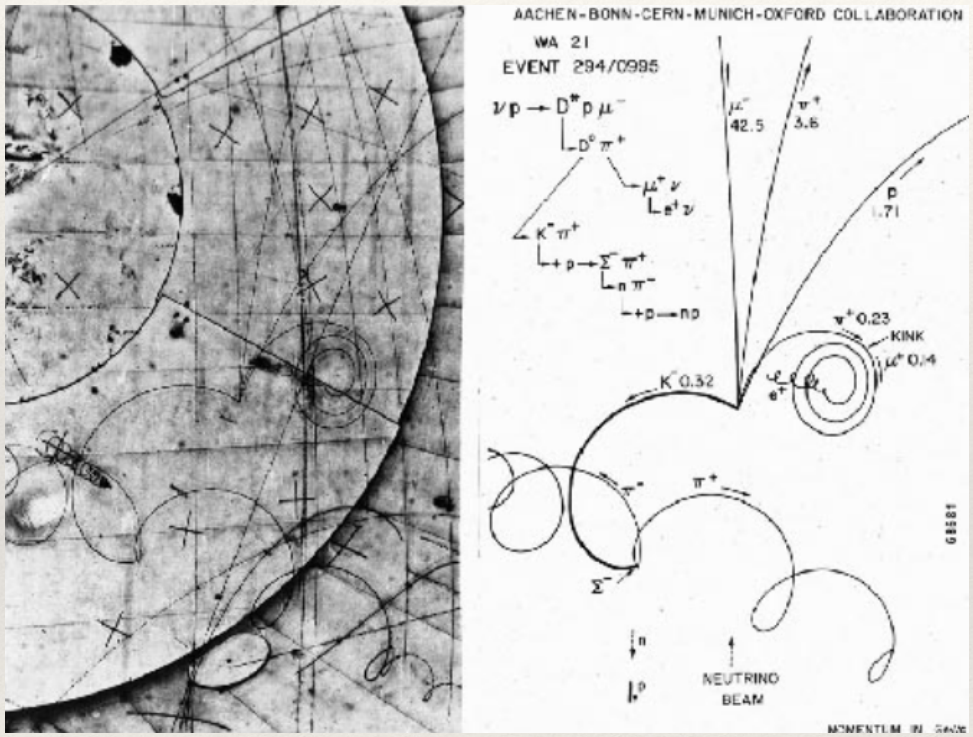
[more at arXiv:1611.05531]

Similar work ongoing at:

- other neutrino experiments - e.g. **NOvA** *[arXiV:1604.01444]*
  - inspired to GoogLeNet architecture. Improvement in the efficiency of selecting electron neutrinos by 40% with no loss in purity. Used as event classifier in both an electron neutrino appearance search, and in a search for sterile neutrinos
- collider experiments in the area of jet physics *[arXiv:1511.05190]* *[arXiv:1603.09349]*

Back to the.. past?

Going beyond feature engineering and embracing the revolutions that DL brings is somehow connecting HEP future to its glorious past..



Neutral currents in BEBC - WA21 CC Charm Event: Roll 204, Frame 995 [CERN]

The data taking pace has changed

- e.g. BEBC in 1973-83 equals to 6 seconds of (e.g.) LHCb today
- e.g. LHC sensor arrays's 1 hr equals to ~ Facebook data in 1 year
- algorithms running on large computing farms took over long ago

Still dealing with inability for humans to visually inspect vast amounts of data

- Indeed, inability "for humans"..

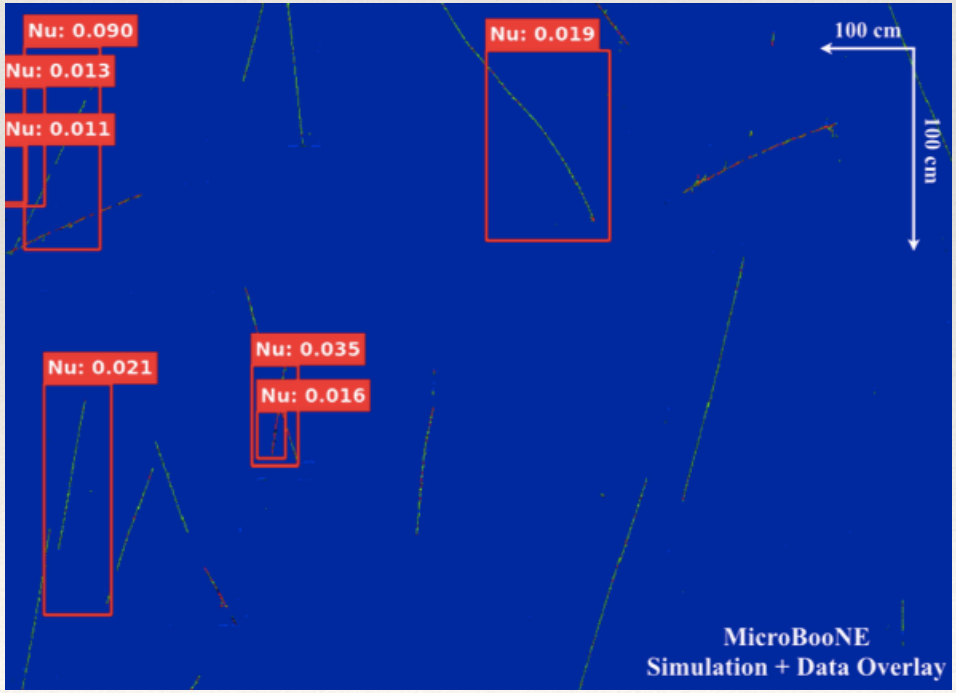
Arguing that "HEP is different"..

Farabet et al. ICML 2012, PAMI 2013

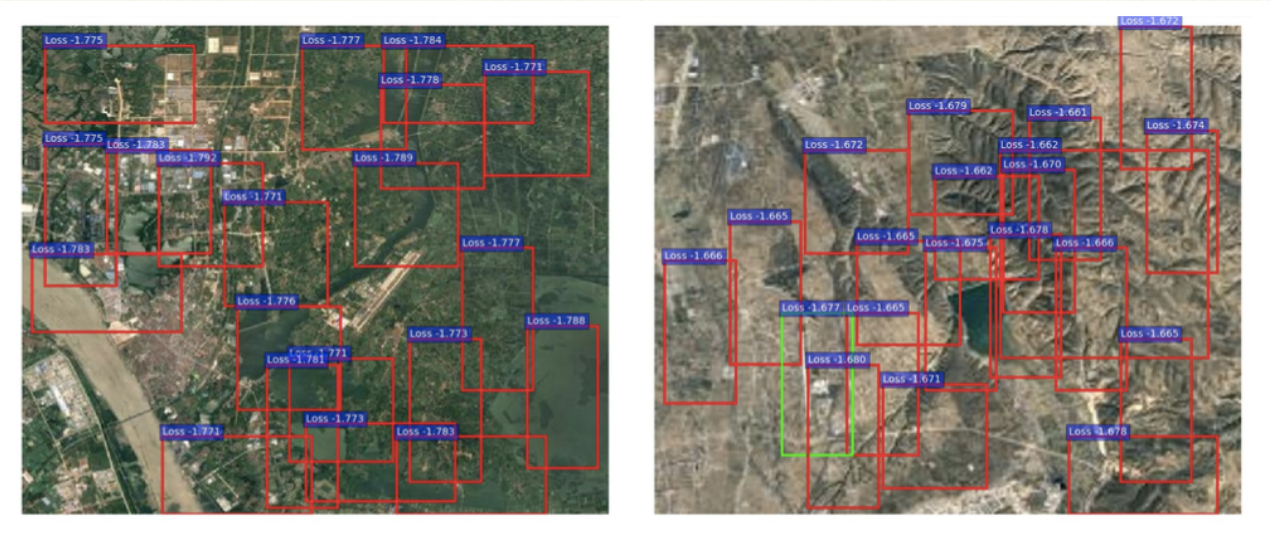
Scene labelling
in automotive applications

MicroBooNE examples of cosmic bkg events with detected neutrino bounding boxes with low scores.

[arXiv:1611.05531]



Arguing that "HEP is different"..

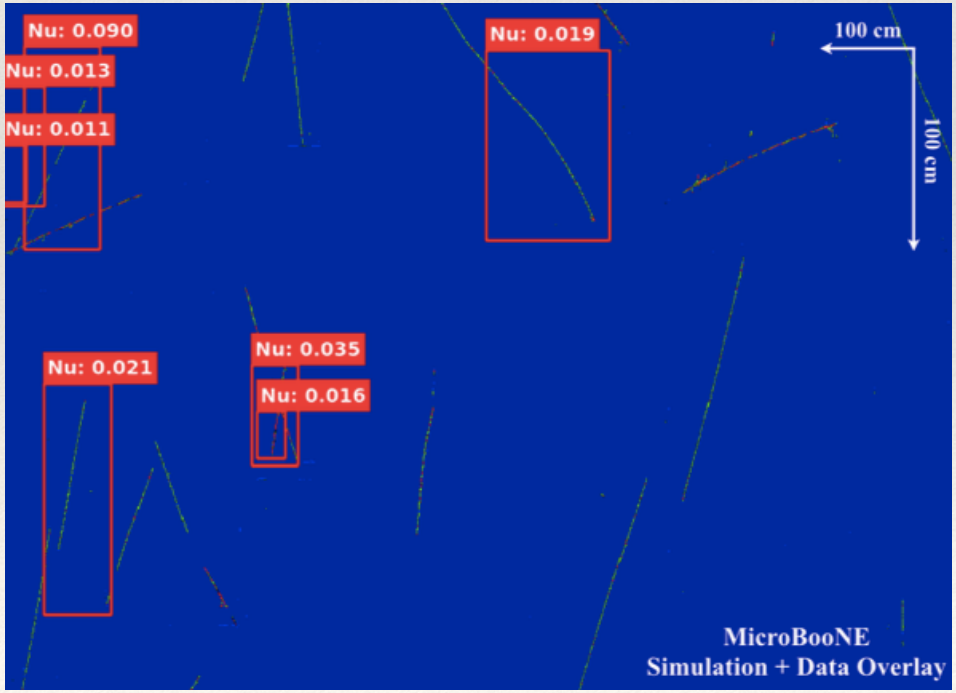


Airports detection from satellite images with CNNs

*[Remote Sens. **2017**, 9, 1198; doi:10.3390/rs9111198]*

MicroBooNE examples of cosmic bkg events with detected neutrino bounding boxes with low scores.

[arXiv:1611.05531]

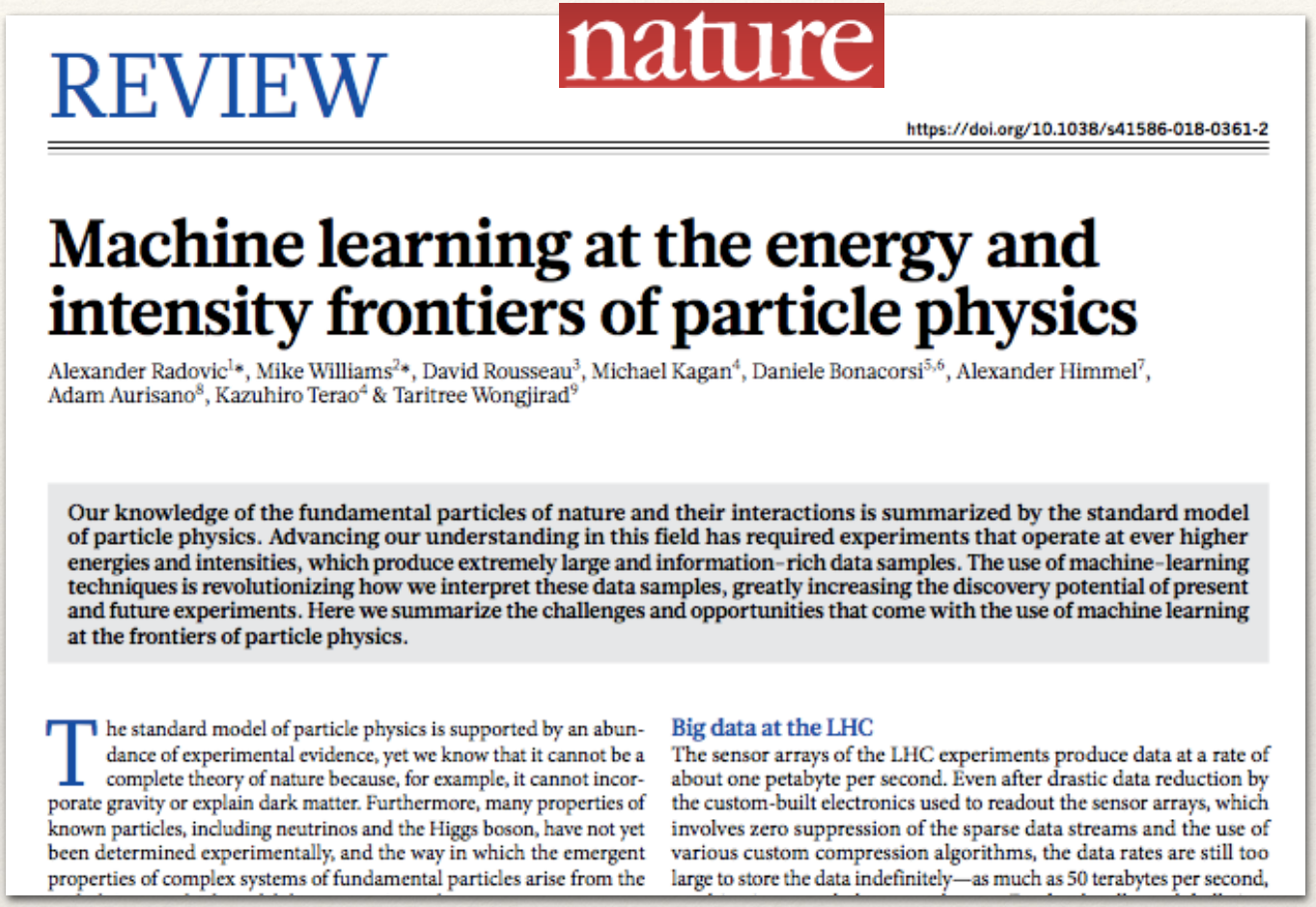


More ML/DL in HEP..

Just the top of the iceberg! More, on a non-exhaustive list, below:

- CNNs/RNNs to reconstruct images from pixel intensities and identify particles and extract many parameters
- Various DL application in the Tracker systems towards High-Luminosity LHC
- Fast generative models like VAEs/GANs as alternatives to HEP Fast Simulation (as Full Simulation is very computationally demanding..), aiming at orders of magnitude improvements!
- unsupervised algos able to monitor many variables at the same time, learn from data and produce an alert when deviations are observed could kill the need of expert shifters in LHC data taking periods
- predictive maintenance studies (algorithms sensitive to subtle signs forewarning of imminent failure, so that pre-emptive actions can be scheduled) on computing centres to reduce the cost of computing while keeping unchanged the physics throughput
- and even more..!
  - hardware-side of choices, deployed computing infrastructures for ML in HEP, jet tagging with RNNs, deep NNs on FPGAs, Deep Kalman Filters, compression using autoencoders, seeking the right format for ML on HEP data, first prototypes of Cloud-compliant ML as-a-service solutions for HEP, ..

More on ML in HEP



REVIEW

nature

https://doi.org/10.1038/s41586-018-0361-2

Machine learning at the energy and intensity frontiers of particle physics

Alexander Radovic[1]*, Mike Williams[2]*, David Rousseau[3], Michael Kagan[4], Daniele Bonacorsi[5,6], Alexander Himmel[7], Adam Aurisano[8], Kazuhiro Terao[4] & Taritree Wongjirad[9]

Our knowledge of the fundamental particles of nature and their interactions is summarized by the standard model of particle physics. Advancing our understanding in this field has required experiments that operate at ever higher energies and intensities, which produce extremely large and information-rich data samples. The use of machine-learning techniques is revolutionizing how we interpret these data samples, greatly increasing the discovery potential of present and future experiments. Here we summarize the challenges and opportunities that come with the use of machine learning at the frontiers of particle physics.

The standard model of particle physics is supported by an abundance of experimental evidence, yet we know that it cannot be a complete theory of nature because, for example, it cannot incorporate gravity or explain dark matter. Furthermore, many properties of known particles, including neutrinos and the Higgs boson, have not yet been determined experimentally, and the way in which the emergent properties of complex systems of fundamental particles arise from the

Big data at the LHC

The sensor arrays of the LHC experiments produce data at a rate of about one petabyte per second. Even after drastic data reduction by the custom-built electronics used to readout the sensor arrays, which involves zero suppression of the sparse data streams and the use of various custom compression algorithms, the data rates are still too large to store the data indefinitely—as much as 50 terabytes per second,

Very recent HEP review work on Nature (Aug 2nd, 2018)

- bit.ly/ML-DBonacorsi

Conclusions

Science is a pool of application areas for ML/DL techniques

- it is an amazing one, indeed

ML-based methods stand as powerful tools in many disciplines

- more recently, DL started to become a game player in some

For the un-initiated, the technology poses significant difficulties

- A constant training path is one of the keys towards success
- **A school like this one is an excellent leg in your ML/DL/DataScience trip!**

Many challenges are just common across sciences..

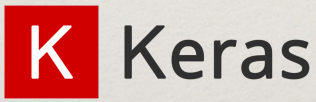
- .. but every science have unique data and tasks, and very peculiar priorities
- and not all sciences are at the same level of advancement and tools adoption

A zoo of always-more-refined algos and techniques to learn!

- more in next slide on this

Frameworks and tools

Regardless of the science (or not!) you focus on, you will go through one (or more) of these:



Availability of world-class ML frameworks is encouraging cross-discipline fertilisation

- scientists from different communities started to talk to each other, and learn from each other's experiences - **like you in this room this week!**
- this might be tough, but will eventually be VERY GOOD.

What should you aim at?!

More similarity in tools/techniques than in applications themselves

- ML/DL/DataScience is like learning a language: it builds bridges to/from other communities

If you e.g. gain experience on one class of algos..

- .. then it will be easier to become expert on neighbouring classes of algos

if you become confident in a ML/DL framework for one application..

- .. then it will be easier to use that experience in other application domains

So, **get started**, **get solid**, and **explore**!

Of course, **Science** welcomes ML/DL practitioners and data scientists, and guarantees that you will never run out of problems to solve!

Thanks for the attention.

Enjoy this week!

References

[H1] https://www.newyorker.com/magazine/2017/04/03/ai-versus-md

[H2] https://arxiv.org/abs/1702.05747

[H3] https://www.nature.com/articles/nature21056

[H4] https://www.biorxiv.org/content/early/2017/05/12/064279.1

[H5] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5355231/

[H6] https://www.prnewswire.com/news-releases/arterys-receives-first-fda-clearance-for-broad-oncology-imaging-suite-with-deep-learning-300599275.html

[IG1] https://arxiv.org/pdf/1406.2661.pdf

[C1] https://pubs.acs.org/doi/abs/10.1021/ci900157k

[C2] https://www.ncbi.nlm.nih.gov/pubmed/21819139

[B1] https://www.ncbi.nlm.nih.gov/pubmed/24860169

[B2] https://academic.oup.com/bib/article/7/1/86/264025

[B3] S. Ananiadou, S. McNaughtJ, "Text Mining for Biology and Biomedicine", Artech House Publishers, Jan 2006

[E1] https://peerj.com/preprints/1720/

[G1] https://ieeexplore.ieee.org/document/8423072/

[G2] https://ieeexplore.ieee.org/document/6729499/

[G3] https://ieeexplore.ieee.org/document/7486265/

[G4] https://arxiv.org/abs/1605.01156

[G5] https://dl.acm.org/citation.cfm?doid=3097983.3098112

[G6] https://arxiv.org/abs/1703.03126