### Introduction to Intelligent systems

D. Bonacorsi (University of Bologna)

Corso CCR Big Data - Bologna, 9-13 September 2019

DISCLAIMER: major focus on (selected) challenges in Science(s).

### Definition(s) of AI/ML/DL

A pictorial definition by a company (Nvidia)



#### Machine Learning

"Classical programming uses rules and data to produce answers. Machine Learning uses data and answers to produce rules."

– F. Chollet

[DISCLAIMER: just to set the context - more discussion on ML next week]



Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

In other words: schools like this one are THE RIGHT PLACE TO BE !

#### Acceleration in the adoption of learning techniques

[ DISCLAIMER: also personal opinions here ]

A revive and acceleration happened recently, mainly because of factors that I would list as:

- the raise of **Big Data** (both as a motivation, and as a helper!)
- the technology progresses (e.g. GPGPUs)
- "Democratisation" [\*] of massive computing resources via Cloud approaches [\*] as most of these resources world-wide are far from being free-of-charge, and Big Data companies implement carefully designed business models, this is debatably "democratic" in a social sense. Not discussed here further, though: here we aim at stating that "in principle" you have a "pay-and-access" option on not-on-premise resources, which did not exist before.

Today, it is a fact that AI/ML/DL are among the core transformative technologies at the basis of most world-wide activities aiming at extracting actionable insight from (Big) Data

#### Al on the Hype cycle (as of July 2019)



### Intelligent systems are ubiquitous

#### 1/ Intelligent Data Processing

Algorithms that <u>automatically</u> analyse specific datasets to <u>extract</u> <u>information</u> that are <u>actionable</u>. E.g.:

 predictive analytics, i.e. study data to provide forecast of a specific phenomenon or process



#### 4 type of Data Analytics

 anomaly detection (e.g. fraud detection), i.e. identification of elements not compatible with a "standard" expected model

#### 2/ Recommendation systems

Exploit the <u>user-provided (directly or indirectly) information</u> (e.g. purchases, selections, ..) to <u>suggest</u> something the user might be interested in.

• e.g. influence the "customer journey" in e-commerce via suggestions to users





#### 3/ Image Processing

Systems able to analyse static images or videos, and detect objects, people, etc.

• e.g. video-surveillance (facial recognition), security



#### 4/ Autonomous Vehicles

Self-driving means of transportation

- more complex and wide that image processing alone
- not only on roads, but ships, drones, planes, ..



#### Example: computer vision





#### 5/ Natural Language Processing

Solutions that process (written or spoken) language

- passive: comprehension and elaboration of text (e.g. translation)
- active: virtual assistants or chatbots, able to activate services upon human trigger via voice (e.g. front-level customer care systems able to perform real-time sentiment analysis, correlations, relevant data extraction, etc)



#### 6/ Intelligent Objects/Robots

Objects able to analyse situations and environment and take autonomous actions with no human intervention whatsoever.

 from "simple" intelligent objects (e.g. suitcase), to "sophisticated" security devices

In industrial processes, <u>Autonomous Robots</u> are able to act just following a basic task description and interaction capabilities (e.g. sensors) with a relatively large set of different environments.

many applications in automation engineering





#### In summary...

In all the aforementioned cases, most common objectives are:

- costs reduction and processes improvement
- increased revenue
- support to decision processes

### Intelligent systems in Science

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" the 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)

### Which sciences?

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, ..)

In the following, focus on <u>highlighting some level of similarity across</u> <u>science challenges</u>, and on <u>how specific ML/DL tools on Big Data might</u> <u>massively help a very diverse set of scientific disciplines</u>

For attendants of this school: **topics covered in this School can make a huge difference in MANY Sciences.** And this is happening already..

DISCLAIMER: I will quote various ML models and NN architectures without going into the details of each.

More later in the week..



<u>Credits</u>: 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

G eoffrey 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 yesno 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

#### [H1]

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. 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 (..) <u>The role of radiologists will evolve from doing</u> <u>perceptual things that could probably be done</u> <u>by a highly trained pigeon to doing far more</u> <u>cognitive things.</u>"

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



### "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 biopsyproven 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 (...)"



#### Dermatologist-level classification of skin cancer with deep neural networks

Andre Esteva<sup>1</sup>\*, Brett Kuprel<sup>1</sup>\*, Roberto A. Novoa<sup>2,3</sup>, Justin Ko<sup>2</sup>, Susan M. Swetter<sup>2,4</sup>, Helen M. Blau<sup>5</sup> & Sebastian Thrun<sup>6</sup>

Skin cancer, the most common human malignancy<sup>1-3</sup>, 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)<sup>4,5</sup> show potential for general and highly variable tasks across many fine-grained object categories<sup>6-11</sup> 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<sup>12</sup>—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<sup>2</sup> working system. The CNN is Trained using 757 disease classes. Our every year. One in five Americans will be diagnosed with a cutaneous dataset is composed of dermatologist-labelled images organized in a malienancy in their lifetime. Although melanomas reversent fewer than tree-structured taxonomy of 2.032 diseases. in which the individual

images (for example, smartphone images) exhibit variability in factors such as zoom, angle and lighting, making classification substantially more challenging<sup>32,34</sup>. We overcome this challenge by using a datadriven 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<sup>16,16,19</sup>, 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<sup>35</sup>, have recently been shown to exceed human performance in visual tasks such as playing Atari games<sup>26</sup>, strategic board games like Go<sup>27</sup> and object recognition<sup>6</sup>. 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<sup>8</sup> that was pretrained on approximately 1.28 million images (1,000 object categories) from the 2014 ImageNet Large Scale Visual Recognition Challenge<sup>6</sup>, and train it on our dataset using transfer learning<sup>28</sup>. 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]

doi:10.1038/nature21056

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] Drop50

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. Metaensemble'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** 

### More: deep generative models [1/2]

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, ...

### More: deep generative models [2/2]

.. 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]

D. Bonacorsi

#### 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<sup>1,2,3,4</sup>, Alexander Aliper<sup>2</sup>, Andrey Kazennov<sup>2,7</sup>, Polina Mamoshina<sup>2,5</sup>, Quentin Vanhaelen<sup>2</sup>, Kuzma Khrabrov<sup>1</sup>, Alex Zhavoronkov<sup>2,6,7</sup>

<sup>1</sup>Search Department, Mail.Ru Group Ltd., Moscow, Russia

<sup>2</sup>Pharmaceutical Artificial Intelligence Department, Insilico Medicine, Inc., Emerging Technology Centers, Johns Hopkins University at Eastern, Baltimore, Maryland, USA

<sup>3</sup>Big Data and Text Analysis Laboratory, Kazan Federal University, Kazan, Republic of Tatarstan, Russia

<sup>4</sup>St. Petersburg Department of V.A. Steklov Institute of Mathematics of the Russian Academy of Sciences, Petersburg, Russia

<sup>5</sup>Department of Computer Science, University of Oxford, Oxford, UK

<sup>6</sup>The Biogerontology Research Foundation, Trevissome Park, Truro TR4 8UN, UK

<sup>7</sup>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 anticancer 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 artificiallyintelligent 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 <u>non</u>-ethical <u>not</u> 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 Corso CCR Big Data - Bologna, 9-13 September 2019

### Outlook on role(s) of ML in Medical Science

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

• ...

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, ..

# 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 expert 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).



#### 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]



Corso CCR Big Data - Bologna, 9-13 September 2019

36

## Bioinformatics



### **Bioinformatics**

"<u>Bioinformatics</u>" = 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



Very complex mapping of ML methods to biological tasks..

### ML in Genomics and Proteomics

#### <u>Genomics</u>

- 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 noncoding 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 similarity (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:

- array 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

#### Microarrays





Genomic20K - Genomic Image of the Core's 20k human Oligonucleotide microarray comparing MCF10 & UACC1179 breast cell line expression profile)

Example of **DNA microarray data**. One measures how much (the colours) certain genes are "expressed" in a group of patients.

### 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 (SVM), 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

## Big (bio)-data and ML/DL in Bioinformatics is a very active and interesting field!

Corso CCR Big Data - Bologna, 9-13 September 2019

44

## Geoscience



#### Geoscience(s)

Earth's major interacting components are **complex dynamic systems** 

• e.g. lithosphere, biosphere, hydrosphere, and atmosphere

Their states <u>perpetually keep changing in space and time</u>, 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..

(Big) **Geo-data** comes mainly from 2 broad categories of sources:

- 1. observational data collected via <u>sensors</u> (space, sea, land)
- 2. <u>simulation</u> 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)
- <u>improvements in computational resources</u> 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 <u>cloud platforms</u>

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 [1/2]

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. <u>advanced</u> 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.

#### Hard to use ML on geo-data [2/2]

#### 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



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

#### follow next talk!

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

### Concluding remarks for science(s)

**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
- Ability to match novel techniques with proper computing infrastructures is the key!
- A school like this one is an excellent leg in your BigData/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 crossdiscipline 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 for your skills portfolio

### What should you aim at?!

More similarity in tools/techniques than in applications themselves

- it 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!

Both Industry and Science welcome BigData/ML/DL workers/ experts, and you are guaranteed that you will never run out of problems to solve!

## Thanks for the attention.

# Enjoy this week!

#### Some references (whenever I did not manage to put them on the slide directly)

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

- [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
- Corso CCR Big Data Bologna, 9-13 September 2019