Deep learning applied to medical image analysis: epistemology and data

Francesca Lizzi

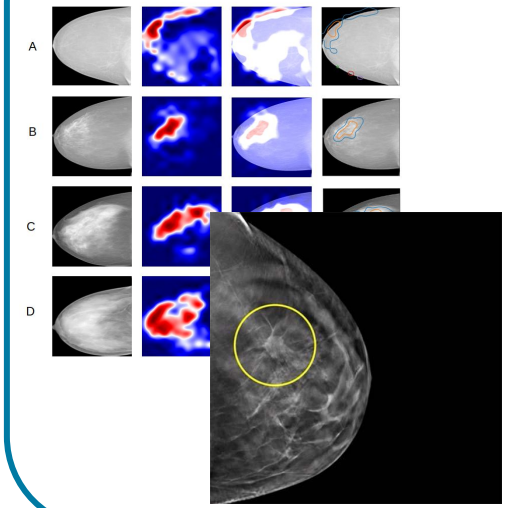
Scuola Normale Superiore, Pisa

National Institute for Nuclear Physics, Pisa

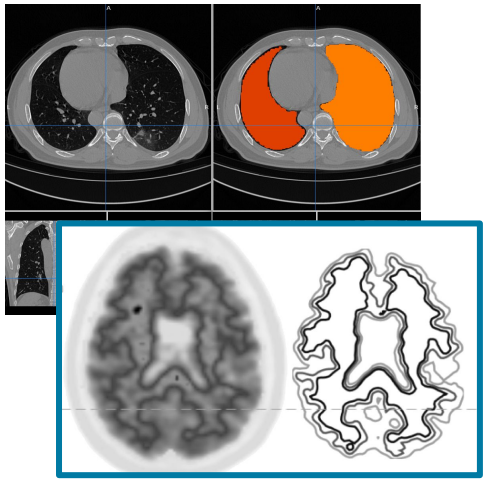
Department of Physics, University of Pisa

Deep Learning and Medical Imaging

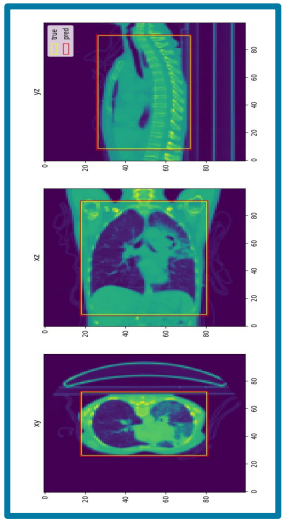
Classification:



Segmentation:



Regression:



- **Super-resolution**: high resolution image reconstruction from low resolution images;
- **Data Generation;**

Changing the scientific paradigm: are we back to 1500?

According to the definition given by Khun in “The Structure of Scientific Revolutions” a paradigm is:

an accepted model or pattern whose fundamental components, for a certain period, remain substantially undisputed

Classical critiques:

- There is only **a little evidence** of this modus operandi.
- Taking into account just a paradigmatic approach produces **too clean and linear stories** on how disciplines evolve, deleting **the pluralism of the history of science.**
- It may f**orce nature to fit into the paradigm.**

Despite the limitations of the definition of paradigm, it is an useful and interesting instrument to delineate a **simplified** history of the **evolution of scientific paradigm.**

Evolution of scientific paradigm: the “Fourth Paradigm”

According to Kitchin and Hey et al., we can delineate a very simplified scheme to classify how scientific paradigms evolved.

Experimental Science:

- Empiricism;
- Describing natural phenomena;

PRE-RENAISSANCE

Theoretical science:

- Modelling;
- Generalization;

PRE-COMPUTERS

Computational science:

- Simulation of complex phenomena;

PRE-BIGDATA

Exploratory science:

- Data-intensive;
- Statistical exploration;
- Data mining

NOW

Rob Kitchin. Big Data, new epistemologies and paradigm shifts. Big Data and Society, 1(1):1–12, 2014.
Tony Hey, Stewart Tansley, and Kristin M. Tolle. Fourth Paradigm, 2021.

Exploratory Science: the “Fourth Paradigm”

Exploratory Science challenges the rules of the traditional scientific paradigm since it is based on the unification of theory, experiment and modelling.

“**Correlation is Enough**”, 2008 Anderson

Data analysts should not propose or even **”bother” themselves with hypothesis** anymore, 2013 Steadman

“Data mining techniques can extract the **complete set of patterns and effects,** producing scientific conclusions without any further experiment.”, 2009 Prensky

Deep learning is “**a kind of magic**”, 2018, Stuart J Russell

Two possible interpretations of the Fourth Paradigm:

- **Inductivist approach:** Big data can capture an entire domain of knowledge; there is no need of a priori model, theory or hypothesis; the application of data mining is agnostic, i.e. **data are inherently meaningful and truthful**; meaning transcends the specific domain.
- **Constructivist approach:** data are **constructed and not natural**; data must contain the information we are looking for (hypothesis); data selection and sampling always introduce a bias; interpretation needs expertise

Are medical images inherently truthful?

Medical images are the results of:

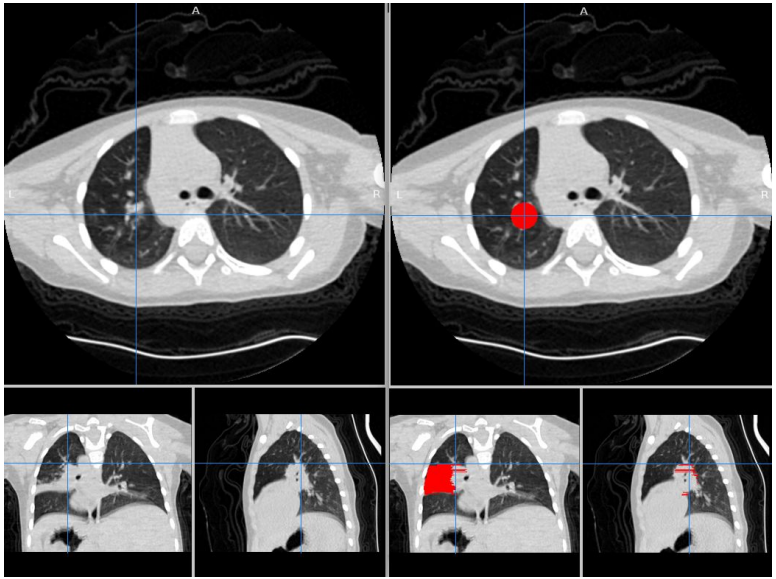
1. A traditional scientific process: physical studies on the interaction between matter and radiation, human body and radiation, on the physical processes of X-ray, magnetic fields and ultrasounds production as regards radiology and radioactivity and all the issues linked to it as regard the nuclear imaging;
2. A technological development history: detectors improvements, the electronics which, simplifying, determine for example important image quality characteristics, reconstruction algorithms and pure technological improvements such as the sliding contacts.
3. An industrial process: medical imaging systems are expensive, there are few vendors that deals with the imaging machinery market. As a result of an industrial process, some parts of the medical images production are protected by patents which inhibits the complete knowledge on how an image is produced.
4. A function-based process: medical images are made on the basis of their utility and improved following their possible uses in hospitals. They are made to be presented to physicians in a way that medical doctors can interpret and taking into account the specific medical formation process they attended. Moreover, contrary to natural images, most of medical imaging modality implies the delivering of a radiation dose to the patients, making their use a dynamic equilibrium between costs and risks, in terms of capital and health, and benefits.

1. Ground Truth on Medical Images

When we train a deep learning algorithm, we always minimize a **loss function** with respect to a **true value**

- GT is made of the opinion of one physician, more than one physician or in a consensus;
- An objective GT may exist: in oncology for example we have the biopsy but it may be necessary to cross patient data taken from different health institutes;
- label noise, polarization of disease spectrum;

Label Noise



CT scan of patient volume-covid19-A-0013
COVID-19 Challenge data set

2. Public and private data

Medical images data sets are scarce and underpopulated

The Cancer Imaging Archive

- 147 collections;
- the most populated is the National Lung Screening Trial (NLST) with **26254** Computed Tomography (CT);
- the second one is the Breast Cancer Screening-DBT, **5060** Digital Breast Tomosynthesis (DBT)

8 datasets contain more than 1000 subjects, 51 datasets between 100 and 1000 and the remaining ones less than 100.

The problem of image format: NIfTI and DICOM: **preservation of information, anonymization and harmonization.**

3. The necessity of a multidisciplinary approach

Statistical Validation -> Intermediate Validation -> Clinical Validation: trial studies -> perspective studies: time consuming and expensive.

Is it true that improving the accuracy of a certain diagnosis, making controlled comparisons between physicians and algorithms, leads to an improvement in clinical practice? Hospitals are COMPLEX ECOSYSTEMS:

Health processes and health organization should be taken into account.

- Example: emergency department could benefit of a fast automated and reliable tool.

We cannot ignore **privacy** policy and management.

Who is responsible for the diagnosis? **Accountability** issues

Why this talk at SIF 2022?

Choosing an epistemological approach that considers **limitations and boundaries is essential** for the development of good quality algorithms.

We should **refuse** the **free hypothesis and theory approach** since it is based on the enhancement of **correlation instead of causation**.

In order to build algorithms that should be used in clinical practice, we have to define the **context** in which they are going to be used.

Context includes **society**, how we frame it, **hospitals policies**, research institutes policy, **funding**, **organization** capability...

What is at stake is our ability to produce knowledge from a critical position, avoiding the accusation of practising ”magic” or ”alchemy”, knowing and assuming how much complex is to create algorithms, especially deep learning one, with the scope of applying them in hospitals and, by assuming it, proceed towards a fair, scientific, active and impacting application of data science to medicine.

Thank you for the attention, questions?

francesca.lizzi@sns.it