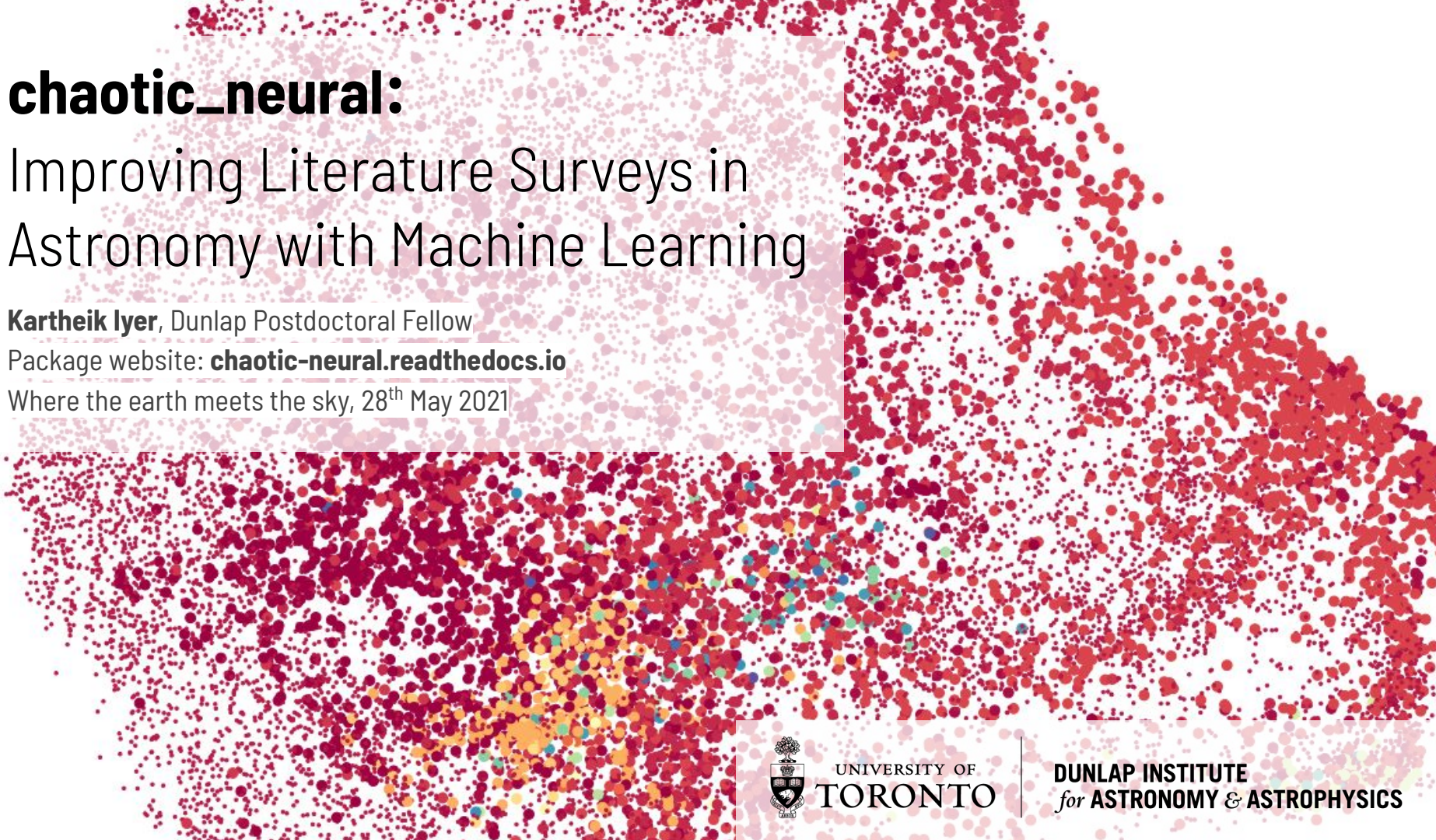
chaotic_neural:

Improving Literature Surveys in Astronomy with Machine Learning

Kartheik Iyer, Dunlap Postdoctoral Fellow

Package website: **chaotic-neural.readthedocs.io**

Where the earth meets the sky, 28th May 2021



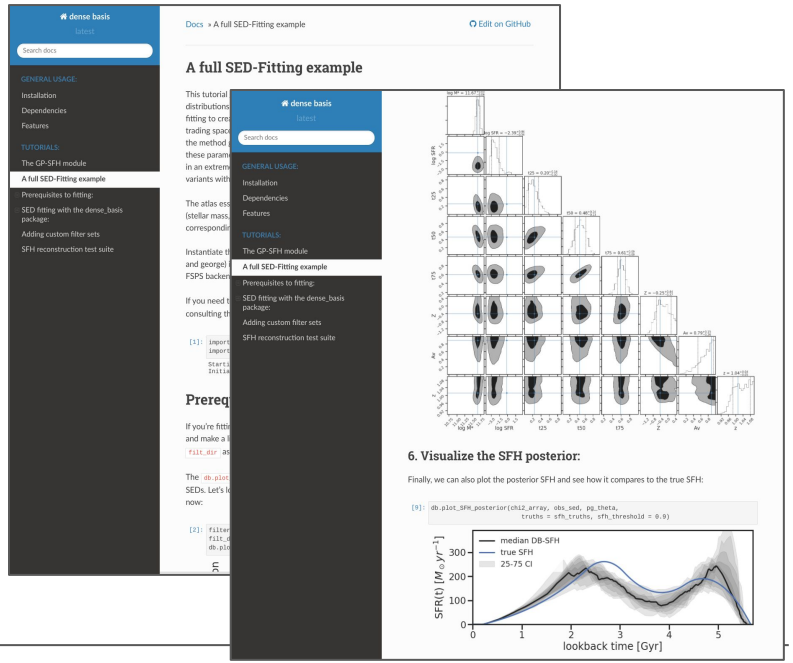
UNIVERSITY OF TORONTO

DUNLAP INSTITUTE *for* ASTRONOMY & ASTROPHYSICS

About me:

I'm **Kartheik Iyer**,
(*pronounced car-thick eye-ear*)
a Dunlap Postdoctoral Fellow at the Dunlap Institute & University of Toronto

I develop tools to interpret observations of distant galaxies,
e.g., the **Dense Basis SED fitting & SFH reconstruction package**
(Iyer et al. 2019, dense-basis.readthedocs.io/
github.com/kartheikiyer/dense_basis)



Broadly, I am interested in
- Galaxy evolution (specifically stochasticity/timescales)
- Reconstructing galaxy star formation histories
- ML applied to astrophysics problems
- Astrostatistics, low S/N problems and Gaussian processes
- Teaching, Mentoring & Physics/Astronomy Outreach

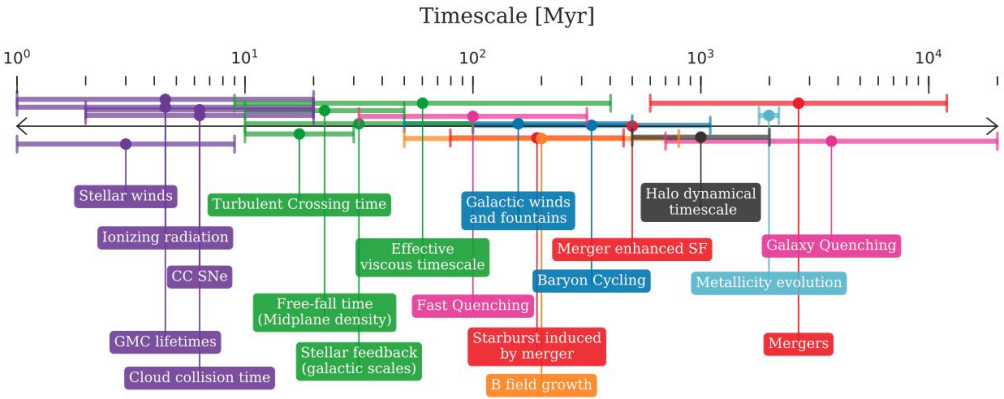
About me:

I'm **Kartheik Iyer**,
(pronounced car-thick eye-ear)
a Dunlap Postdoctoral Fellow at the Dunlap Institute & University of Toronto

I develop tools to interpret observations of distant galaxies,
e.g., the **Dense Basis SED fitting & SFH reconstruction package**
(Iyer et al. 2019, dense-basis.readthedocs.io/
github.com/kartheikiyer/dense_basis)

Broadly, I am interested in
- Galaxy evolution (specifically stochasticity/timescales)
- Reconstructing galaxy star formation histories
- ML applied to astrophysics problems
- Low S/N problems and Gaussian processes
- Teaching, Mentoring & Physics/Astronomy Outreach

Figure: Timescales in galaxy evolution (Iyer et al. 2020)



chaotic_neural: the status quo

In astronomy, searching for literature has many avenues:

- the SAO/NASA **Astrophysics Data System** (**ADS**):
https://ui.adsabs.harvard.edu/
(has a whole bunch of visualisation and other tools!)
- the **arxiv**.org preprint server:
https://arxiv.org/
- **google scholar**:
https://scholar.google.com/
(recommendations based on what you add to your favourites)
- **arxivsorter**:
https://arxivsorter.org/papers/new
(FoF sorter for ArXiv papers)
- **peer/community recommendations**:
e.g. journal clubs, group meetings, benty-fields etc.

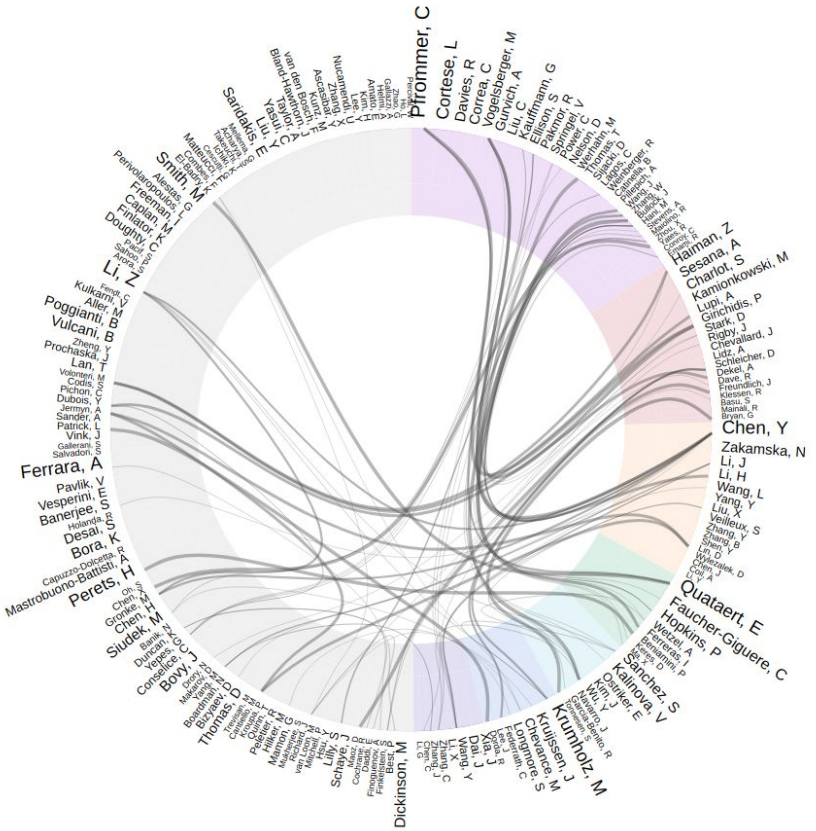


Figure: ADS author network figure for the query 'galaxy evolution' [link; built using 1000 most recent papers]

chaotic_neural: the status quo (contd.)

In astronomy, searching for literature has many avenues:

- the SAO/NASA **Astrophysics Data System** (**ADS**):
https://ui.adsabs.harvard.edu/
(has a whole bunch of visualisation and other tools!)
- the **arxiv**.org preprint server:
https://arxiv.org/
- **google scholar**:
https://scholar.google.com/
(recommendations based on what you add to your favourites)
- **arxivsorter**:
https://arxivsorter.org/papers/new
(FoF sorter for ArXiv papers)
- **peer/community recommendations**:
e.g. journal clubs, group meetings, benty-fields etc.

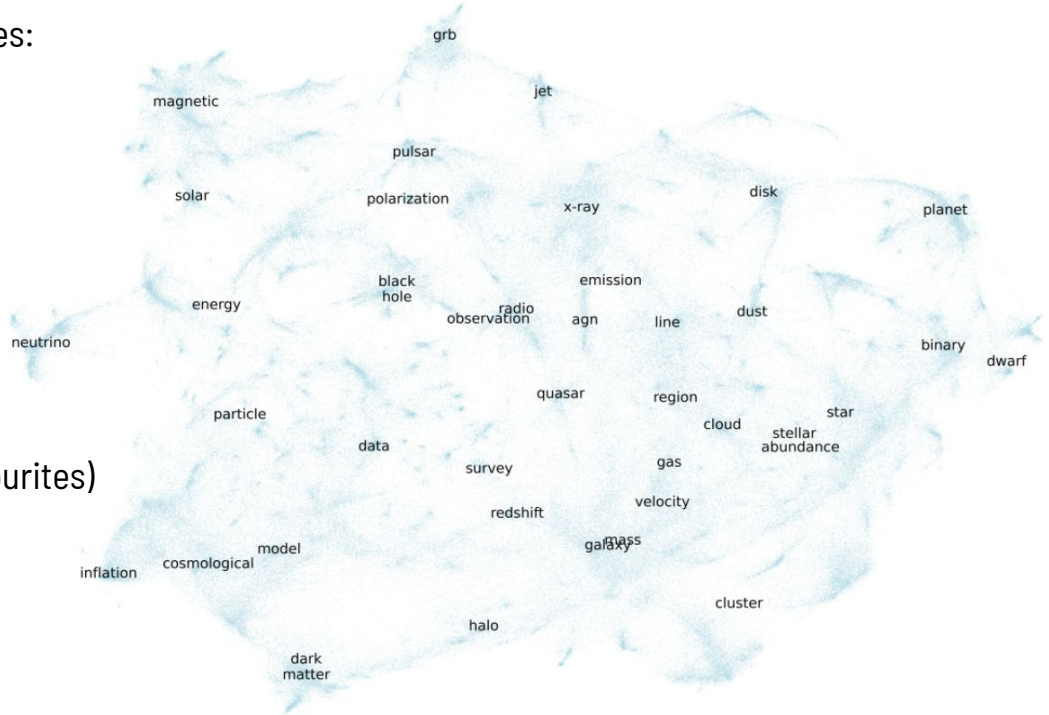


Figure: Arxivsorter map showing astrophysics paper distribution [link; Credits: Manuchehr Taghizadeh-Popp & Brice Ménard]

chaotic_neural: the need for a better model

However, there remain a number of issues:

- **large volume of new literature / day**
(makes it difficult to keep up with emerging trends & recent work)

- **specialized jargon**
(makes it difficult to find and search for the right keywords)
(e.g. SFR-M correlation, Star-Forming Sequence (SFS) and Star-Formation Main Sequence (SFMS) all refer to the same observed correlation between the stellar masses and star formation rates of galaxies)*

- **in-group bias toward referencing & citing papers**
(makes it difficult to find papers outside your collaboration network)

- **meta-analysis is difficult**
(contradictory results for similar phenomena are often the result of differing methodology, data reduction, sample size, or sample-selection in datasets)

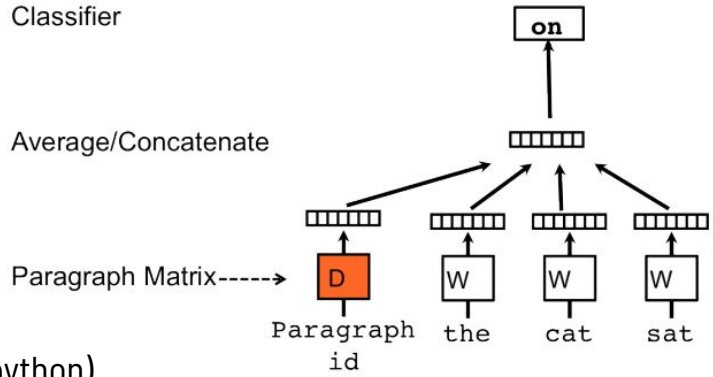
chaotic_neural: context-aware literature searches

A potential solution:

- (train an ML method to) **search for and classify papers by context**.
  & **find similar papers based off a starting paper** (or a list of keywords).

- many possible methods exist (transformers, VAEs, etc.)
  as well as many metrics (continuous bag-of-words, skip-gram
  latent dirichlet allocation, etc.)

- trying something interesting here: **Doc2Vec** (Le & Mikilov 2014; gensim in python)
  generalization of earlier word2vec algorithm



> Train a network to learn a mapping:
> ***paper abstract -> vector in high-dim space***
> in a way that encodes context

Letter | Published: 03 July 2019

Unsupervised word embeddings capture latent knowledge from materials science literature

Vahe Tshitoyan, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova, Kristin A. Persson, Gerbrand Ceder & Anubhav Jain

Nature **571**, 95–98 (2019) | Cite this article

- use learned feature vectors corresponding to paper abstracts for all kinds of things!
(e.g. discoveries in material science - see abstract above & corresponding mat2vec git repo)

(this is very much an emerging field)

A Survey of Deep Learning for Scientific Discovery

Maithra Raghu[1,2*] Eric Schmidt[1,3]
[1] Google
[2] Cornell University
[3] Schmidt Futures

Abstract

Over the past few years, we have seen fundamental breakthroughs in core problems in machine learning, largely driven by advances in deep neural networks. At the same time, the amount of data collected in a wide array of scientific domains is dramatically increasing in both size and complexity. Taken together, this suggests many exciting opportunities for deep learning applications in scientific settings. But a significant challenge to this is simply knowing where to start. The sheer breadth and diversity of different deep learning techniques makes it difficult to determine what scientific problems might be most amenable to these methods, or which specific combination of methods might offer the most promising first approach. In this survey, we focus on addressing this central issue, providing an overview of many widely used deep learning models, spanning visual, sequential and graph structured data, associated tasks and different training methods, along with techniques to use deep learning with less data and better interpret these complex models — two central considerations for many scientific use cases. We also include overviews of the full design process, implementation tips, and links to a plethora of tutorials, research summaries and open-sourced deep learning pipelines and pretrained models, developed by the community. We hope that this survey will help accelerate the use of deep learning across different scientific domains.

REVIEW

ADVANCED ENERGY MATERIALS
www.advenergymat.de

A Critical Review of Machine Learning of Energy Materials

Chi Chen, Yunxing Zuo, Weike Ye, Xiangguo Li, Zhi Deng, and Shyue Ping Ong**

Machine learning (ML) is rapidly revolutionizing many fields and is starting to change landscapes for physics and chemistry. With its ability to solve complex tasks autonomously, ML is being exploited as a radically new way to help find material correlations, understand materials chemistry, and accelerate the discovery of materials. Here, an in-depth review of the application of ML to energy materials, including rechargeable alkali-ion batteries, photovoltaics, catalysts, thermoelectrics, piezoelectrics, and superconductors, is presented. A conceptual framework is first provided for ML in materials science, with a broad overview of different ML techniques as well as best practices. This is followed by a critical discussion of how ML is applied in energy materials. This review is concluded with the perspectives on major challenges and opportunities in this exciting field.

1. Introduction

Machine learning (ML) is the branch of artificial intelligence that deals with the development of algorithms and models that can automatically learn patterns from data and perform tasks without explicit instructions. While ML models and algorithms have been known since the 1950s, it is only in the recent decade that the systematic generation and curation of data on unprecedented scales—coupled with exponential increases in computational costs and poor scaling still limit their effectiveness in exploring unconstrained chemical spaces and/or complex real-world materials. For instance, high-throughput DFT screening works typically limit the search space to hundreds or, at best, thousands of materials, while DFT simulations of materials are mostly limited to typically less than 1000 atoms, i.e., bulk crystals and isolated molecules. ML therefore offers a solution to the materials exploration problem, making predictions of new materials or properties from existing data, which in turn can drive the generation of more data that can be used to further refine the ML models.

Here, we will provide an in-depth, critical review of ML-guided design and discovery of energy materials, a field where a nove
energy
have a
climate
typical
Figure
of ML

materials in silico,[19–22]

Unsupervised Manifold Clustering of Topological Phononics

Yang Long, Jie Ren,* and Hong Chen
Center for Phononics and Thermal Energy Science, China-EU Joint Center for Nanophononics, Shanghai Key Laboratory of Special Artificial Microstructure Materials and Technology, School of Physics Sciences and Engineering, Tongji University, Shanghai 200092, China

(Received 1 August 2019; accepted 13 April 2020; published 6 May 2020)

Classification of topological phononics is challenging due to the lack of universal topological invariants and the randomness of structure patterns. Here, we show the unsupervised manifold learning for clustering topological phononics without any *a priori* knowledge, neither topological invariants nor supervised trainings, even when systems are imperfect or disordered. This is achieved by exploiting the real-space projection operator about finite phononic lattices to describe the correlation between oscillators. We exemplify the efficient unsupervised manifold clustering in typical phononic systems, including a one-dimensional Su-Schrieffer-Heeger–type phononic chain with random couplings, amorphous phononic topological insulators, higher-order phononic topological states, and a non-Hermitian phononic chain with random dissipations. The results would inspire more efforts on applications of unsupervised machine learning for topological phononic devices and beyond.

DOI: 10.1103/PhysRevLett.124.185501

Deep learning for fabrication and maturation of 3D bioprinted tissues and organs

Wei Long Ng [a,b]*, Alvin Chan[c]*, Yew Soon Ong[c] and Chee Kai Chua[d]

[a]Singapore Centre for 3D Printing (SC3DP), School of Mechanical and Aerospace Engineering, Nanyang Technological University (NTU), Singapore; [b]HP-NTU Digital Manufacturing Corporate Lab, Singapore; [c]School of Computer Science and Engineering, Nanyang Technological University (NTU), Singapore; [d]Engineering Product Development (EPD) Pillar, Singapore University of Technology and Design (SUTD), Singapore

ABSTRACT

Bioprinting is a relatively new and promising tissue engineering approach to solve the problem of donor shortage for organ transplantation. It is a highly-advanced biofabrication system that enables the printing of materials in the form of biomaterials, living cells and growth factors in a layer-by-layer manner to manufacture 3D tissue-engineered constructs. The current workflow involves a myriad of manufacturing complexities, from medical image processing to optimisation of printing parameters and refinements during post-printing tissue maturation. Deep learning is a powerful machine learning technique that has fuelled remarkable progress in image and language applications over the past decade. In this perspective paper, we highlight the integration of deep learning into 3D bioprinting technology and the implementation of practical guidelines. We address potential adoptions of deep learning into various 3D bioprinting processes such as image-processing and segmentation, optimisation and *in-situ* correction of printing parameters and lastly refinement of the tissue maturation process. Finally, we discuss implications that deep learning has on the adoption and regulation of 3D bioprinting. The synergistic interactions among the field of biology, material and deep learning-enabled computational design will eventually facilitate the fabrication of biomimetic patient-specific tissues/organs, making 3D bioprinting of tissues/organs an impending reality.

ARTICLE HISTORY
Received 22 March 2020
Accepted 16 May 2020

KEYWORDS
3D bioprinting; 3D printing; biofabrication; deep learning; machine learning

chaotic_neural: *let's get to it!*

Github repo:
https://github.com/kartheikiyer/chaotic_neural/

Documentation:
https://chaotic-neural.readthedocs.io/en/latest/

Conference notebook (google colab):
http://bit.ly/chaotic_neural_workbook

Interested in contributing?
email: kartheik.iyer@dunlap.utoronto.ca