Towards High-End Scalability on Bio-Inspired Computational Models

Darío Dematties George K. Thiruvathukal Silvio Rizzi

CYBERCOLOMBIA
THIRD HPC SUMMER SCHOOL: BIO & DATA SCIENCE
2020

CURRENT AI

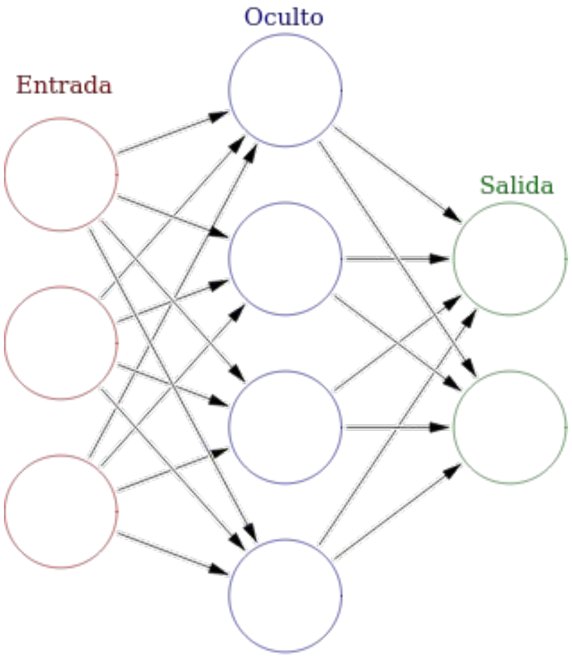


Image source:
https://es.wikipedia.org/wiki/Archivo:Colored_neural_network_es.svg

BIOLOGICAL NEURONS

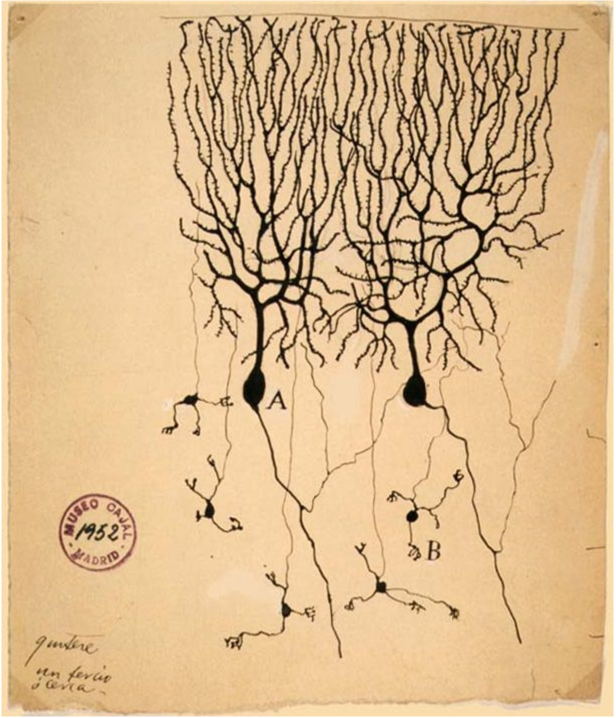


Image source:
https://commons.wikimedia.org/wiki/File:PurkinjeCell.jpg#filelinks

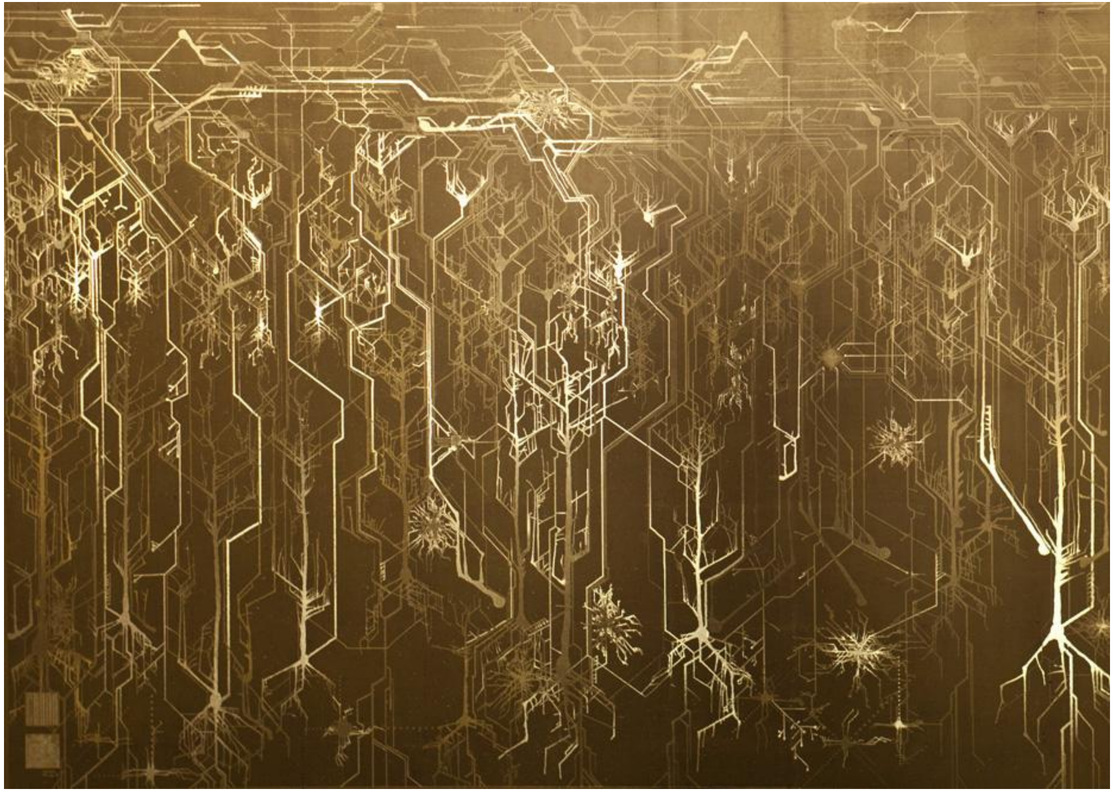


Image credit: Greg Dunn and Brian Edwards

What level of detail is necessary to mimic the neocortex complexity?

- We do not want to mimic all the biologically inherited complexity of the brain.
- Backpropagate or not?
- How to feedback cost functions and what those cost functions should be.

Who is working on this?

NUMENTA
ON INTELLIGENCE

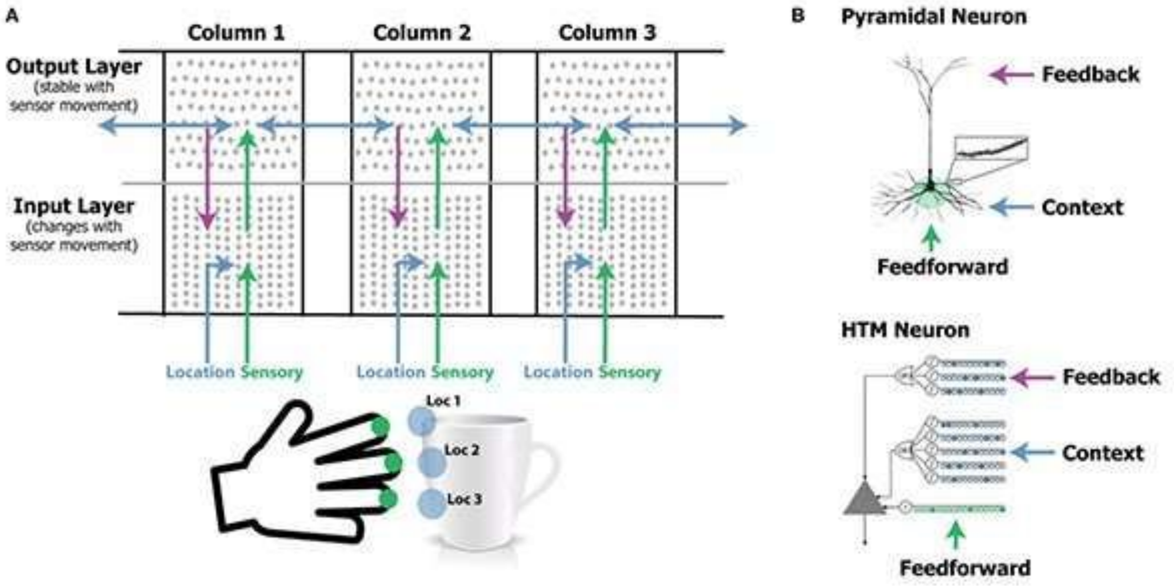


Image credit Numenta Inc. Source:
https://www.google.com/url?sa=i&url=https%3A%2F%2Fmedicalxpress.com%2Fnews%2F2017-11-theory-brain-sensations-mental.html&psig=AOvVaw2DaKNIUV8zFm2nE-2xaqtx&ust=1595423767164000&source=images&cd=vfe&ved=0CAIQjRxqFwoTCIipzPu23uoCFQAAAAAdAAAAABAD

Who is working on this?

cortical.io

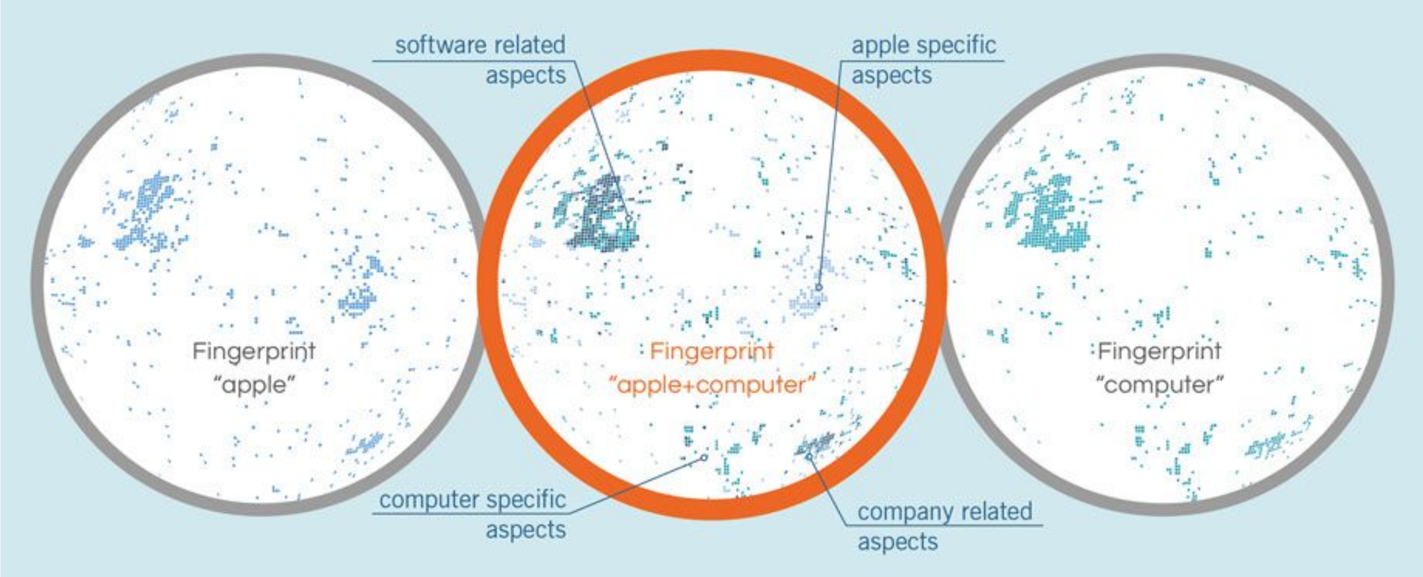


Image credit Cortical io. Source:
https://www.google.com/url?sa=i&url=https%3A%2F%2Faiparis.fr%2F2017%2Fdesousa.html&psig=AOvVaw11slcQ2fwr7fvxcZo6rbQH&ust=1595424125686000&source=images&cd=vfe&ved=0CAIQjRxqFwoTCJCa9Zq43uoCFQAAAAAdAAAAABAD

Darío Dematties

Postdoctoral Researcher at Instituto de Ciencias Humanas, Sociales y Ambientales, CCT CONICET Mendoza

ddematties@mendoza-conicet.gob.ar

George K. Thiruvathukal

Professor of Computer Science, Loyola University Chicago and Visiting Researcher at Argonne National Laboratory

gkt@cs.luc.edu

Silvio Rizzi

Assistant Computer Scientist at the Argonne Leadership Computing Facility

srizzi@alcf.anl.gov

LOYOLA UNIVERSITY CHICAGO

FACULTAD DE INGENIE
Universidad de Bueno

Argonne NATIONAL LABORATORY

Part 1

- Cortical / Sparse / Dendritic thinking
- NLP Applications
  - Phonetics and Grammar
- Brief look at how to understand our model
- What’s in Part 2?
- Q&A / Break

Cortical / Sparse / Dendritic thinking

Activation Sparsity (Ahmad and Hawkins, 2015)

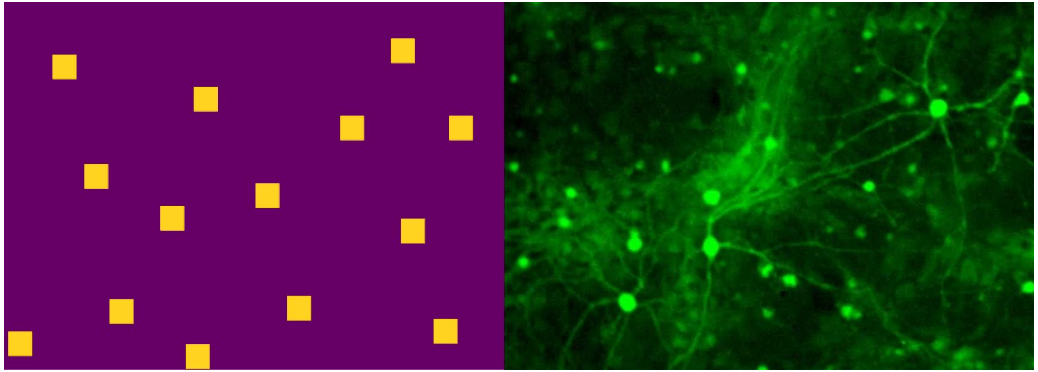
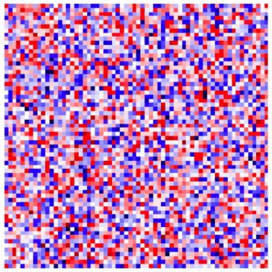


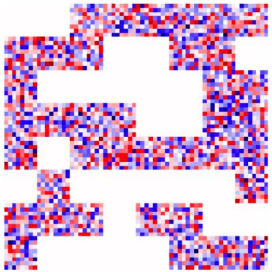
Image from https://www.youtube.com/watch?v=tRPuVAVXk2M

- We simulate Activation Sparsity by means of Sparse Distributed Representations (SDRs).
- SDRs have powerful mathematical properties.

Related work by OpenAI: GPU Kernels for Block-Sparse Weights.



Dense weights



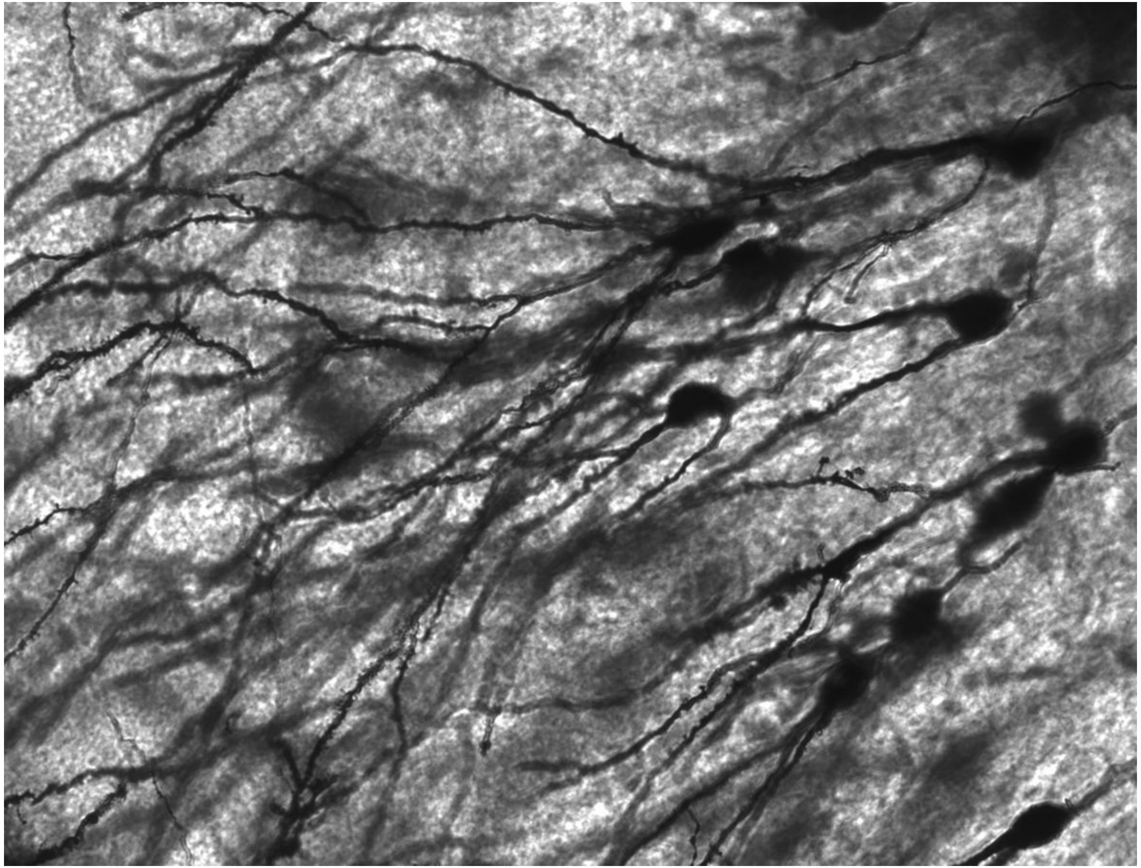
Block-sparse weights

0	0	1	1	1	1	1	1
1	1	1	0	0	1	1	0
1	0	0	0	0	0	1	1
1	1	1	1	0	0	1	1
1	0	1	1	1	1	1	1
0	1	0	0	0	0	0	0
1	1	1	0	1	1	0	0
1	0	0	0	0	1	1	1

Corresponding sparsity pattern

Available at:
https://openai.com/blog/block-sparse-gpu-kernels/

- Highly-optimized GPU kernels for networks with block-sparse weights.
- Sparsity at the block level.
- Each block is densely connected.
- In our work we lead sparsity to the CC level and inside CCs as well.



Activation and Connectivity Sparsity.

Dendritic Compartmentalization

Science

SHARE

REPORT

Dendritic action potentials and computation in human layer 2/3 cortical neurons

Albert Gidon[1], Timothy Adam Zolnik[1], Pawel Fidzinski[2,3], Felix Bolduan[4], Athanasia Papoutsi[5], Panayiota Poirazi[5], Martin H...
+ See all authors and affiliations

Science 03 Jan 2020:
Vol. 367, Issue 6473, pp. 83-87
DOI: 10.1126/science.aax6239

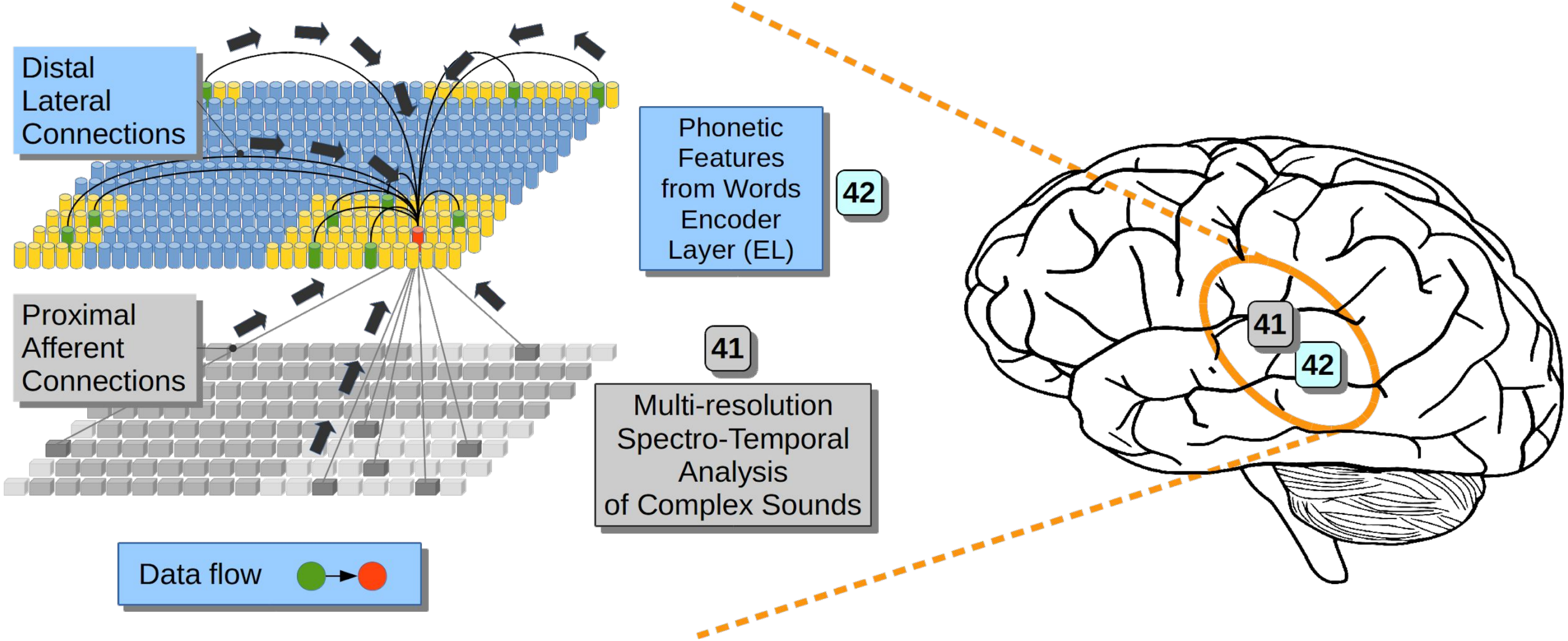
Human dendrites are special

A special developmental program in the human brain drives the disproportionate thickening of cortical layer 2/3. This suggests that the expansion of layer 2/3, along with its numerous neurons and their large dendrites, may contribute to what makes us human. Gidon *et al.* thus investigated the dendritic physiology of layer 2/3 pyramidal neurons in slices taken from surgically resected brain tissue in epilepsy patients. Dual somatodendritic recordings revealed previously unknown classes of action potentials in the dendrites of these neurons, which make their activity far more complex than has been previously thought. These action potentials allow single neurons to solve two long-standing computational problems in neuroscience that were considered to require multilayer neural networks.

NLP Applications

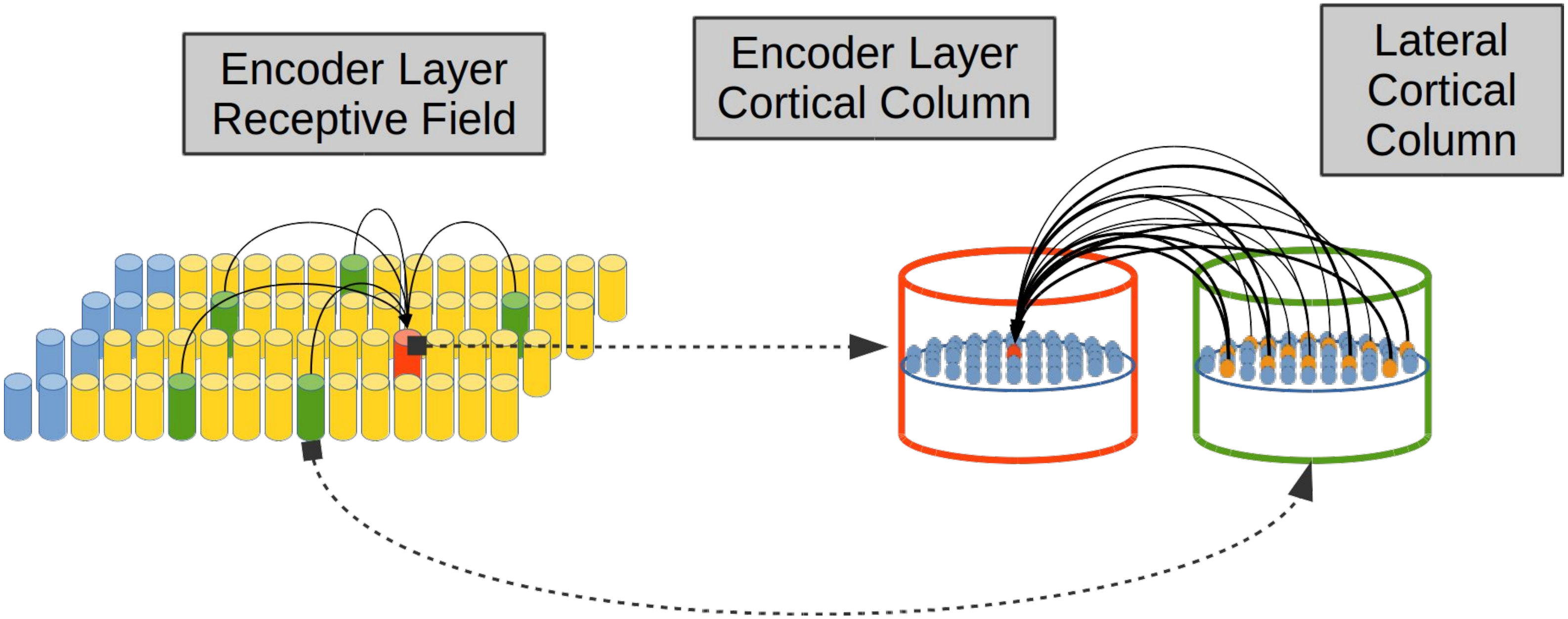
Phonetics

High Level Phonetic Features for Word Discrimination



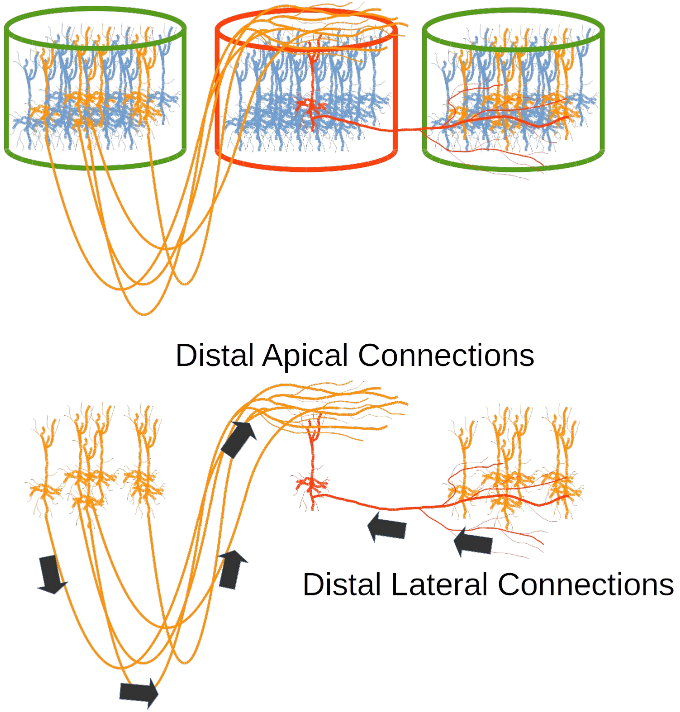
Adapted from: CC0 1.0 Universal (CC0 1.0) Public Domain Dedication. https://svgsilh.com/image/155655.html
Imagen adaptada desde https://doi.org/10.1371/journal.pone.0217966 bajo licencia CC-BY.

Inter and Intra-Columnar Connectivity



Adapted from https://doi.org/10.1371/journal.pone.0217966 under CC-BY licence.

Pyramidal Neurons in Cortical Layer 2/3



Adapted from https://doi.org/10.1371/journal.pone.0217966 under CC-BY licence.

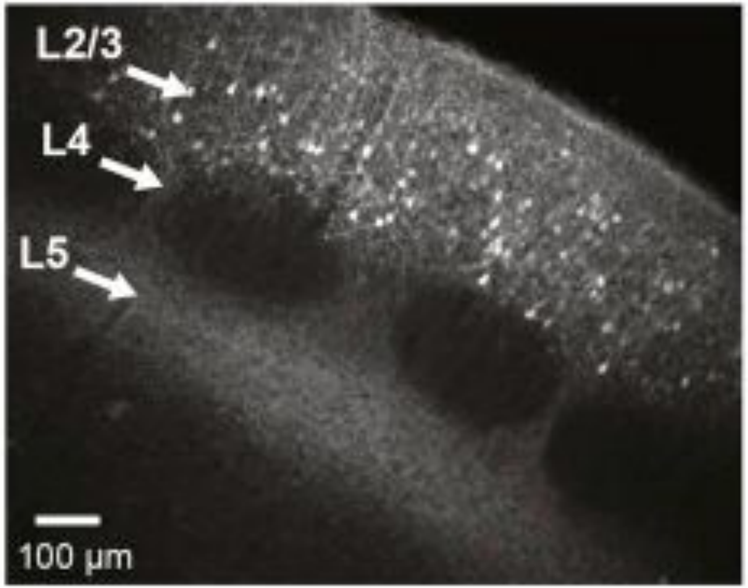
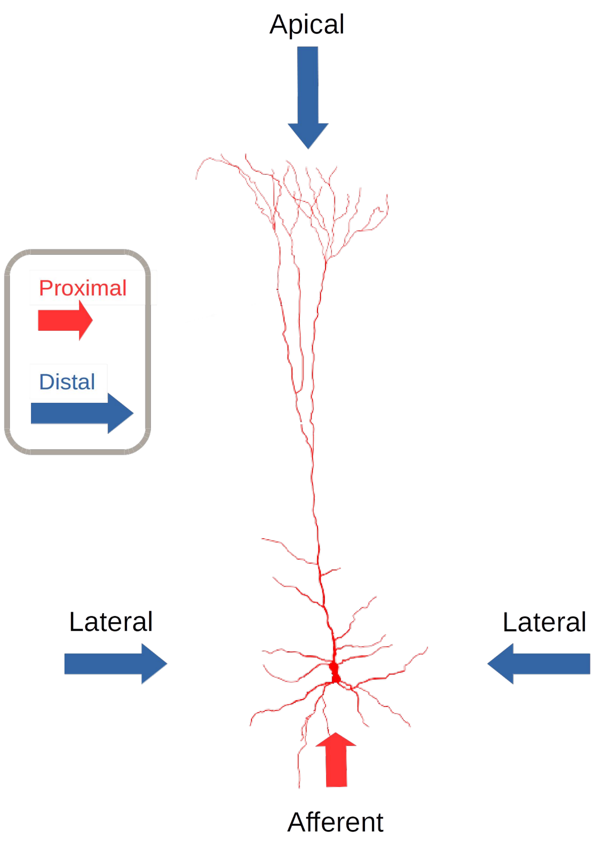
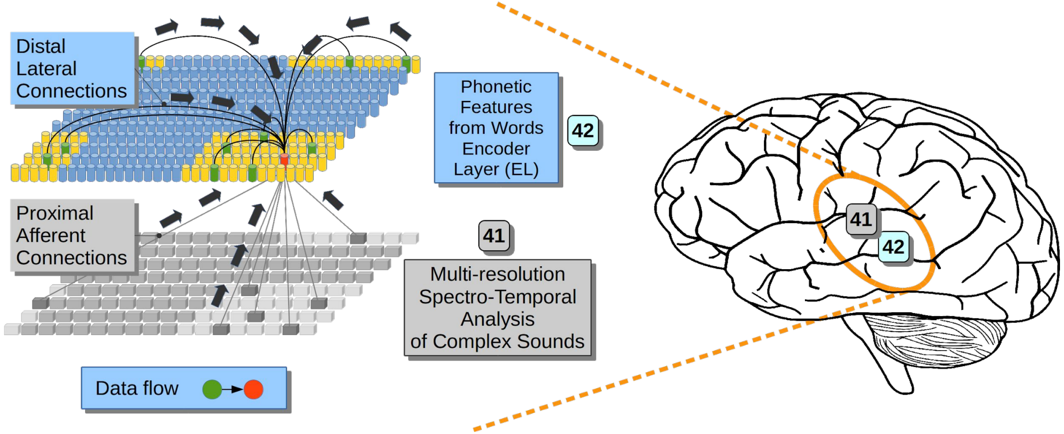
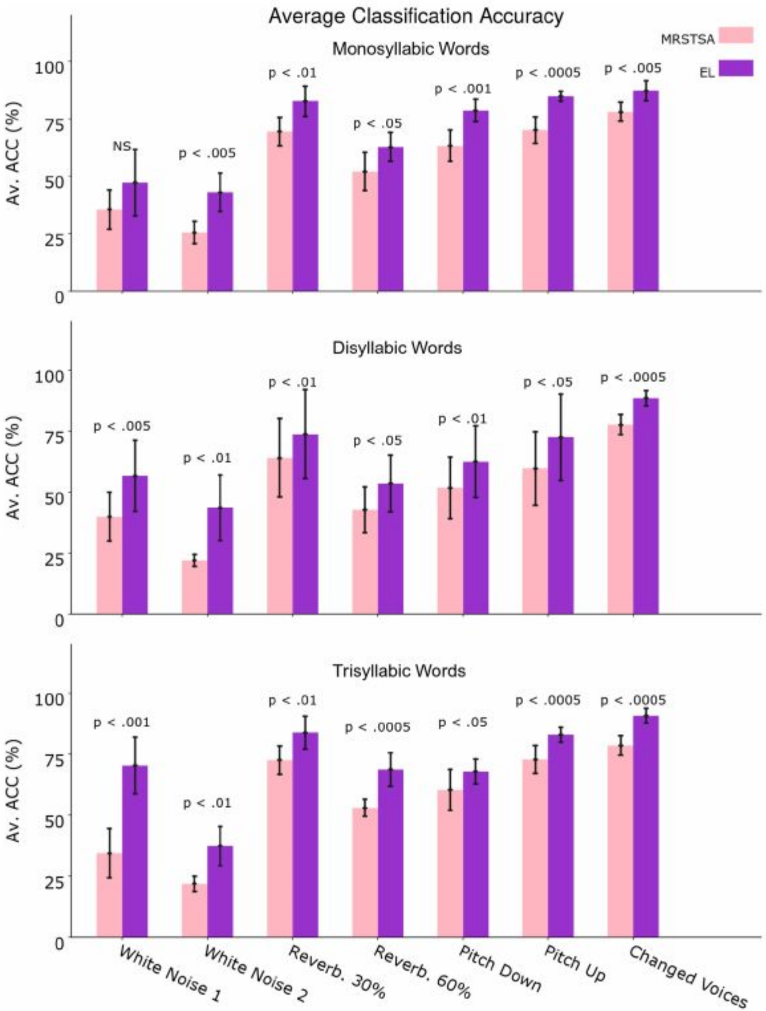


Image source: https://www.sciencedirect.com/science/article/pii/S2211124718313093#fig1 under Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) licence.

Dendritic Compartmentalization in Pyramidal Neurons





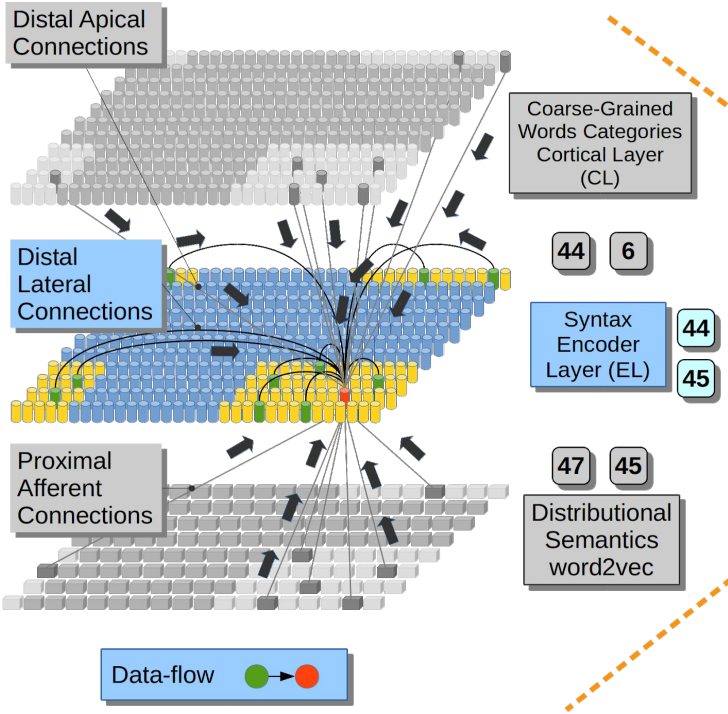


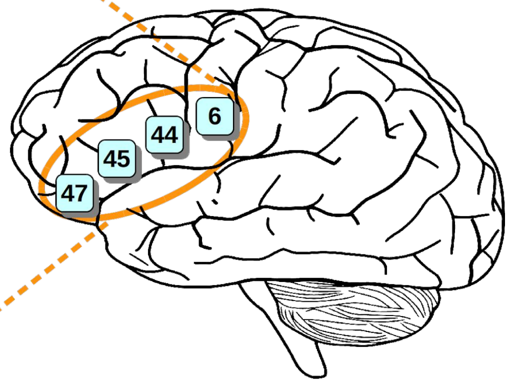
Classification Performance in front of different Acoustic Variants for mono, di and trisyllabic words

The Encoder Layer outperforms the MRSTSA for all the experimental conditions

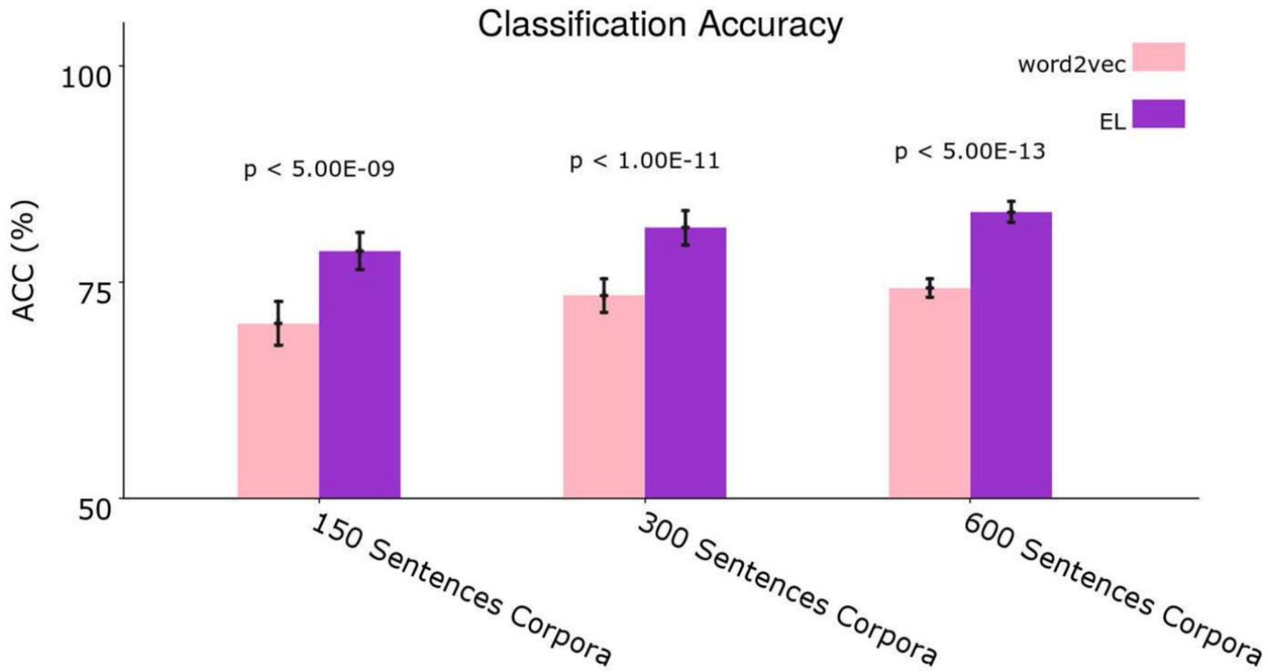
Grammar

Words Grammar Features Acquisition in Sentences





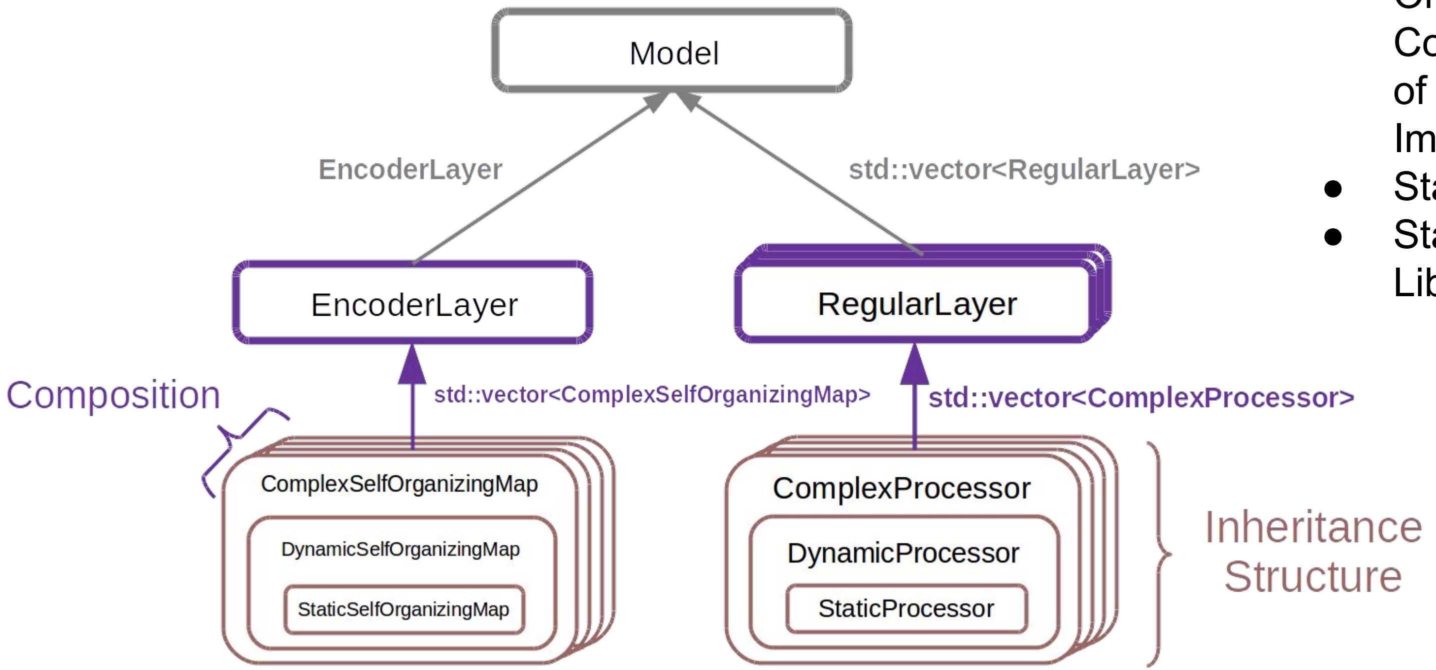
Inspired in the Linguistic Gradient Found in the Left Inferior Frontal Gyrus



The Encoder Layer Outperformed word2vec for all experimental conditions

Image adapted from https://doi.org/10.1371/journal.pone.0217966 under CC-BY licence.

Hierarchical Inheritance and Compositional Structure of the Implementation



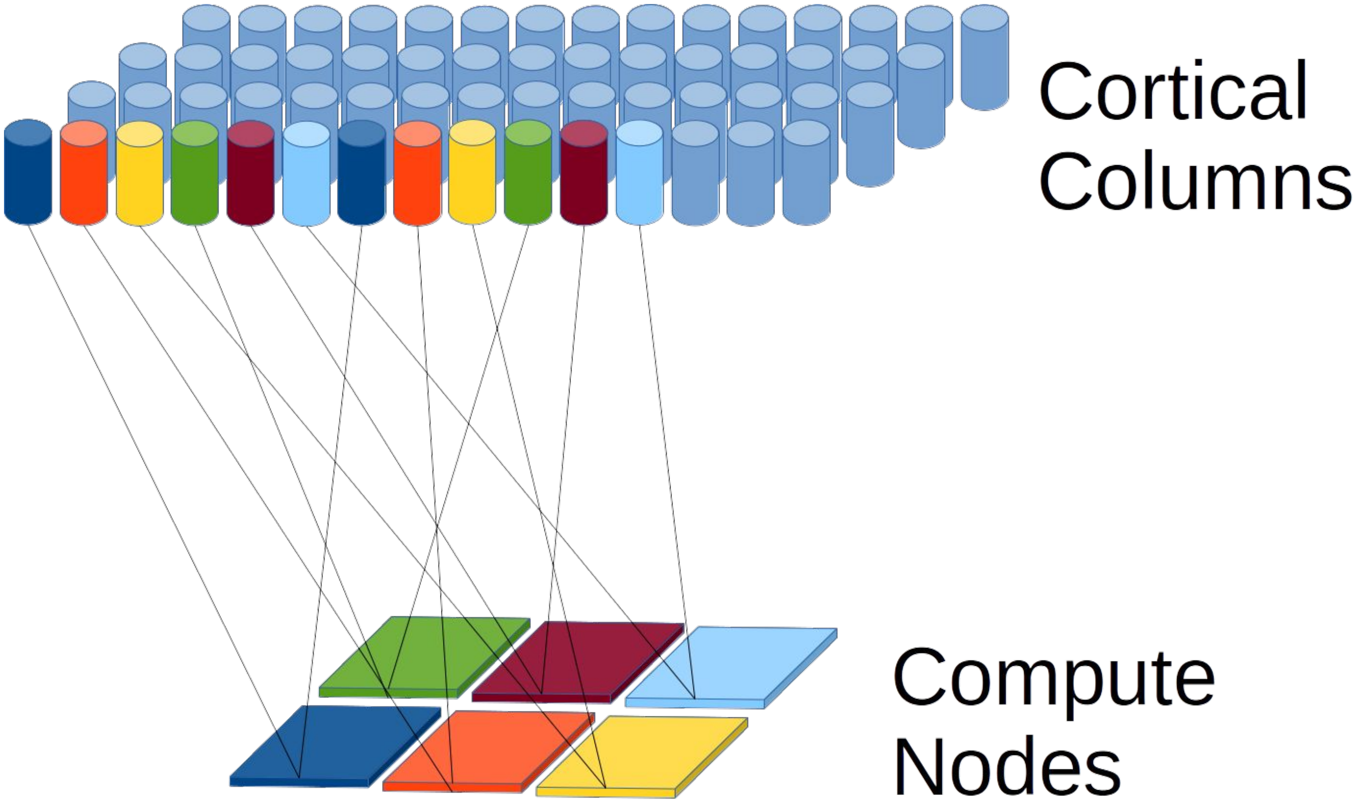
- This is the Object Oriented Inheritance and Compositional Structure of our Model Implementation
- Standard C++ 14
- Standard Template Libraries (STLs)

What’s Next?

Part 2

- Why do we need HPC?
- Hybrid MPI+OpenMP
- Performance results (Parallel Computing Conference)
- Running on Cooley (video)
- Initial explorations / experiences on Theta
- A look at our testbed (Cooley, Theta)
- What’s in Part 3?
- Q&A

Why do we need HPC?



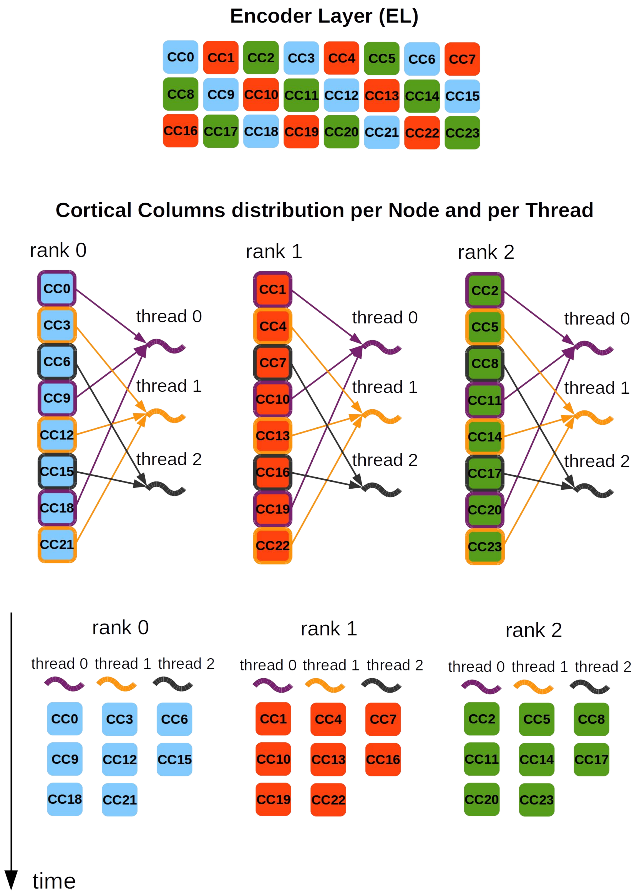


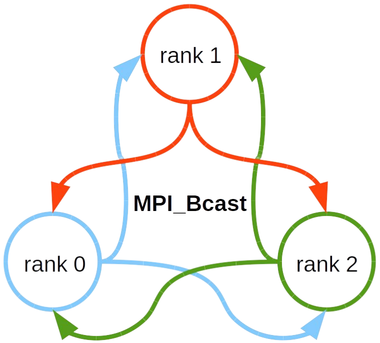
Image adapted from http://ebooks.iospress.nl/volumearticle/53956 under CC BY-NC 4.0 licence.

Towards a High Scalability in Bio-Inspired Models

- MPI + OpenMP
  - MPI: Distributed Memory, distributes CCs in the EL per compute node.
  - OpenMP: Shared Memory, distributes CCs inside a node among different running threads.
- No SIMD (Single Instruction, Multiple Data)
  - This is tipically used by GPUs.
- Coalescence
- Connectivity Randomness and Sparsity

Message Passing Scheme

Inter-Process Communication



Communication Protocol in each MPI rank

std::vector<std::size_t> activeUnits

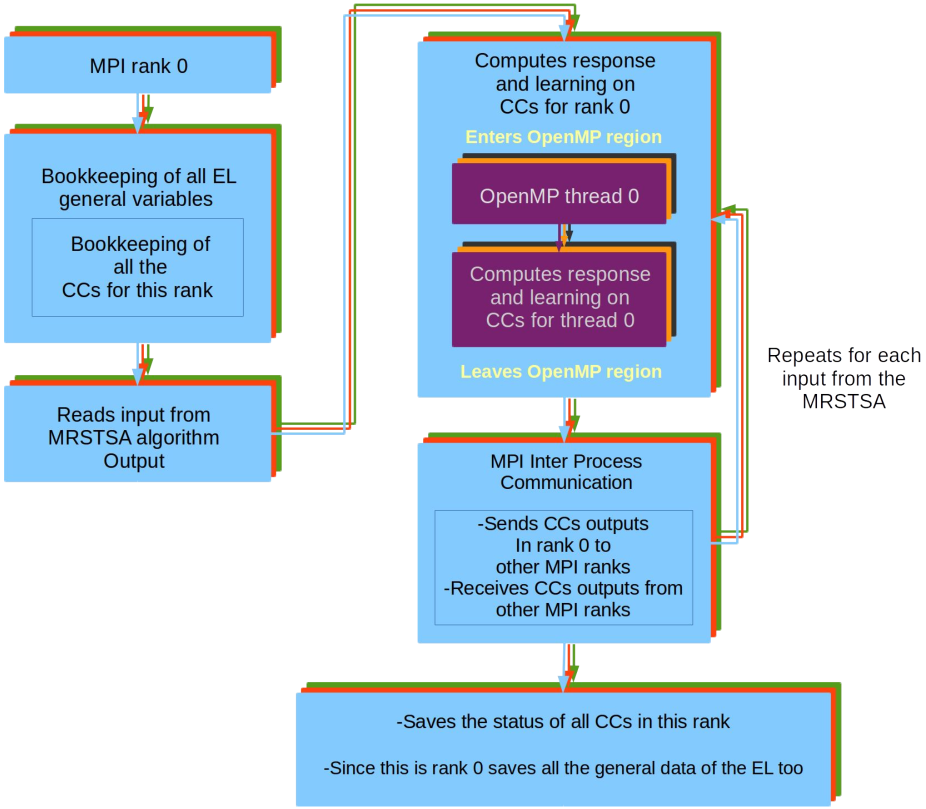
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14		n-3	n-2	n-1	n
15	3	1	10	2	0	67	88	29	133	44	12	23	1	11	…	43	4	27	312

std::vector<std::size_t> boundaries

0	1	2	3		m
4	3	5	2	…	3

- Information among MPI ranks must be transferred in each time step.
- Each MPI rank has to call MPI Bcast just once in order to transmit its data.
- Special communication protocol

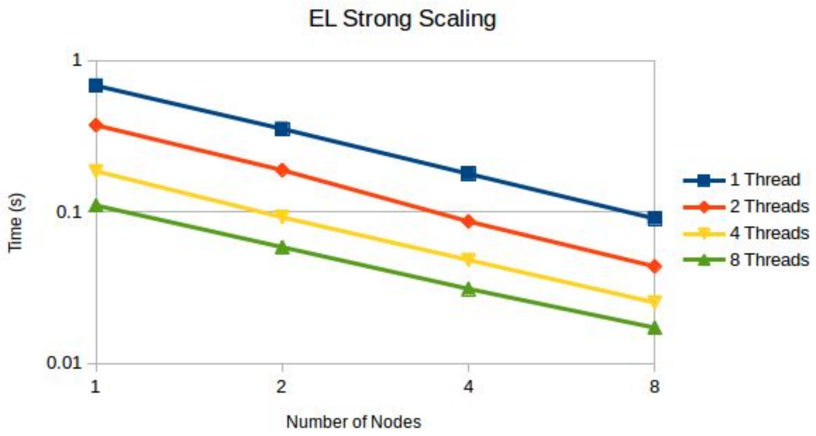
Parallelism Scheme

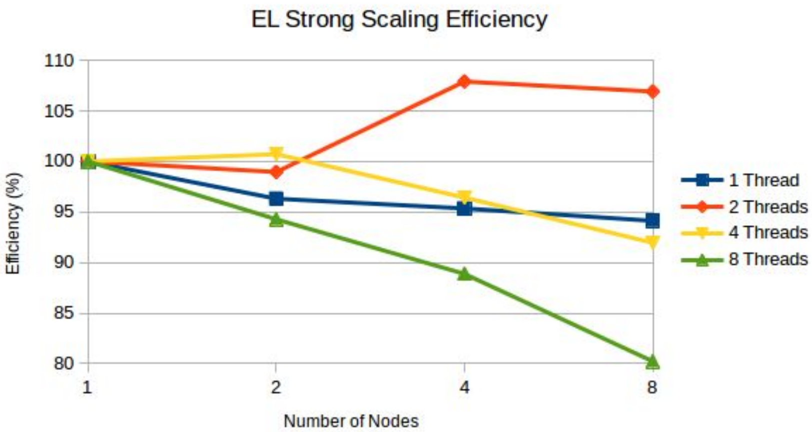


- Each MPI rank keeps a private copy of the entire Network structure (called EL).
- But Each MPI rank keeps only the data of the Cortical Columns (CCs) which corresponds to it.
- Each MPI rank loads a private copy of the inputs that come from the input.
- Then each MPI rank processes the input information by means of only the CCs which are under its charge.

Strong Scaling Results

$$t_1/(N * t_N) * 100$$

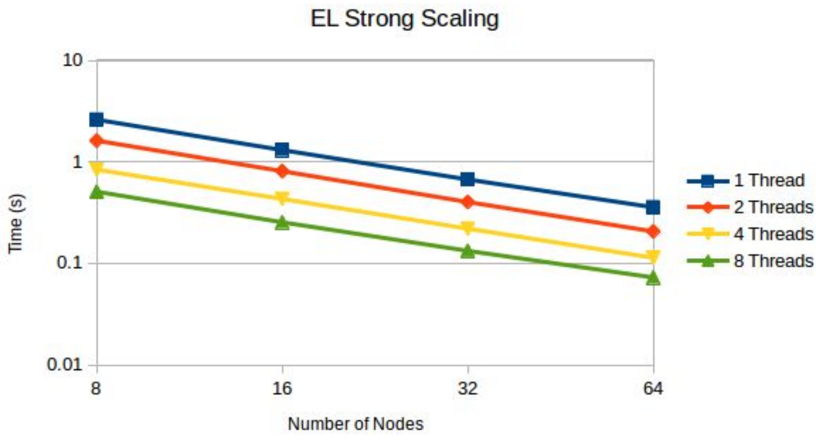


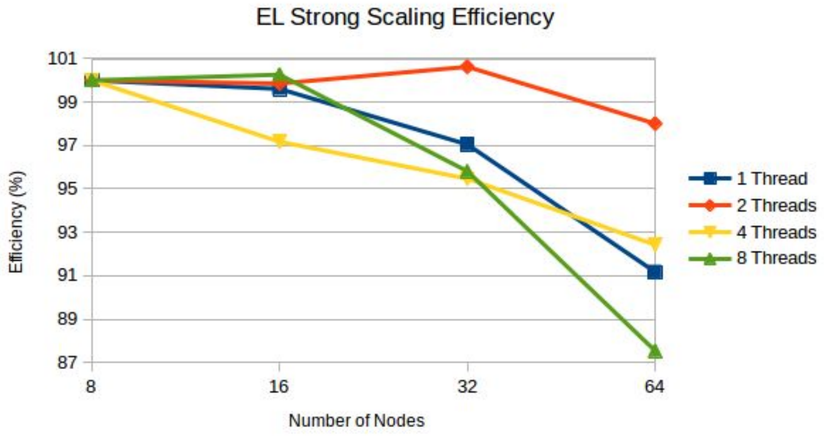


Model	Number of CCs	Afferent RF	Afferent %	Lateral RF	Lateral %	Population Dimensionality	Potential %
Normal	23 x 23	5 x 127	5%	9 x 9	90%	15 x 15	3%

Strong Scaling Results

$$t_1/(N * t_N) * 100$$

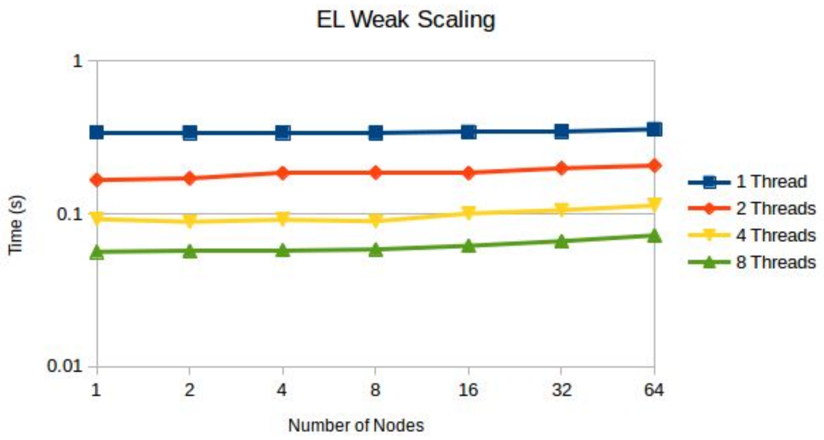


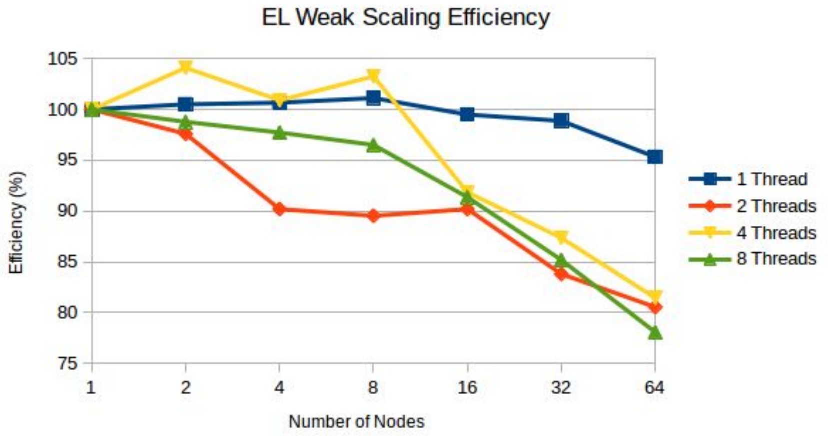


Model	Number of CCs	Afferent RF	Afferent %	Lateral RF	Lateral %	Population Dimensionality	Potential %
Big	128 x 128	5 x 127	5%	9 x 9	90%	15 x 15	3%

Weak Scaling Results

$$t_1/t_N * 100$$





Number of Nodes	Number of CCs	Afferent RF	Afferent %	Lateral RF	Lateral %	Population Dimensionality	Potential %
1 Node	16 x 16	5 x 127	5%	9 x 9	90%	15 x 15	3%
2 Nodes	16 x 32	5 x 127	5%	9 x 9	90%	15 x 15	3%
4 Nodes	32 x 32	5 x 127	5%	9 x 9	90%	15 x 15	3%
8 Nodes	32 x 64	5 x 127	5%	9 x 9	90%	15 x 15	3%
16 Nodes	64 x 64	5 x 127	5%	9 x 9	90%	15 x 15	3%
32 Nodes	64 x 128	5 x 127	5%	9 x 9	90%	15 x 15	3%
64 Nodes	128 x 128	5 x 127	5%	9 x 9	90%	15 x 15	3%

Building the code

- Code available at https://github.com/neurophon/neurophon
- Dependencies: compiler with support for C++14, MPI, HDF5
- Build with `make`
- Video showing the entire process:
    - https://anl.box.com/s/lt3szc36p76b0z7ezmjoakxybu531aq3

Generating a model

- Generate a model. Edit `octave/GenerateModelFiles.m`
- From Octave, run `GenerateModelFiles("Semantic_Model_Aux")`. A directory will be created with three .mat files
- Download data from Zenodo https://zenodo.org/record/2576130#.Xx-Z45NKgiV
- Run `run_Semantic_Model_AUX.sh`
- Video showing the entire process:
  - https://anl.box.com/s/ftrko1o75x4zdg48hp1xiokgh6f388zv

Running in parallel with MPI

- Prepare submission script
- Submit with `qsub`
- Monitor the queue
- Inspect results available in standard output of the run
- Video showing the entire process:
    - https://anl.box.com/s/19kwmz1vhdokxm5tssb14lr7eooaj74g

Cooley: Analytics/Visualization cluster

Peak 223 TF

126 nodes; each node has

- Two Intel Xeon E5-2620 Haswell 2.4 GHz 6-core processors
- NVIDIA Tesla K80 graphics processing unit (24GB)
- 384 GB of RAM

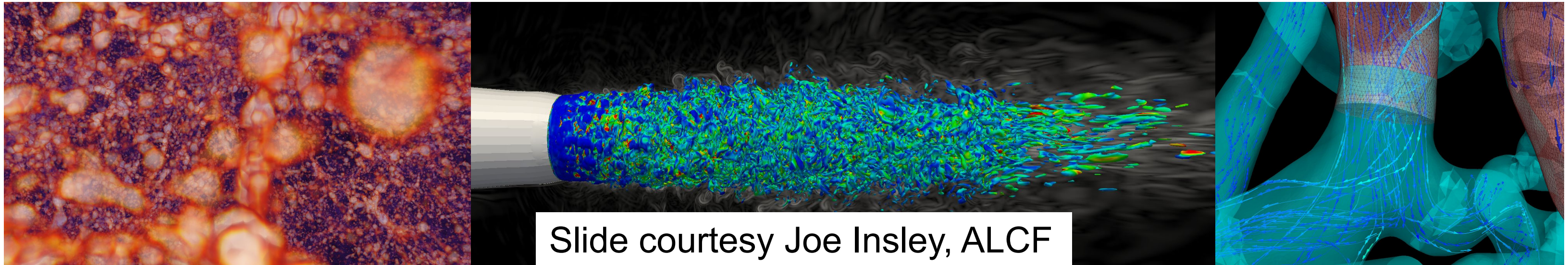
Aggregate RAM of 47 TB

Aggregate GPU memory of ~3TB

Cray CS System

216 port FDR IB switch with uplinks to our QDR infrastructure

Mounts the Theta file system



Slide courtesy Joe Insley, ALCF

Computing Resource for 2020

Theta Cray XC40
4,392 nodes
281,088 cores
892 TiB RAM
Peak flop rate: 11.69 PF

Iota Intel/Cray XC40
44 nodes
2,816 cores
8.9 TiB RAM
Peak flop rate: 117 TF

Firestone IBM Power8
2 nodes + K80 GPU
20 cores
128 GB RAM
Hybrid CPU/GPU

Cooley Cray/NVIDIA
126 nodes
1512 Intel Haswell CPU cores
126 NVIDIA Tesla K80 GPUs
48 TB RAM / 3 TB GPU

Storage Capability

Disk

- Theta: ~18 PB of GPFS/Lustre file system capacity; 9PB is GPFS and 9.2PB is Lustre.

Tape

- The ALCF has three 10,000-slot libraries using LTO 6 tape technology. The LTO tape drives have built-in hardware compression for an effective capacity of 36-60 PB.

Theta

Features Intel processors and interconnect technology, a new memory architecture, and a Lustre-based parallel filesystem – all integrated by Cray's HPC software stack

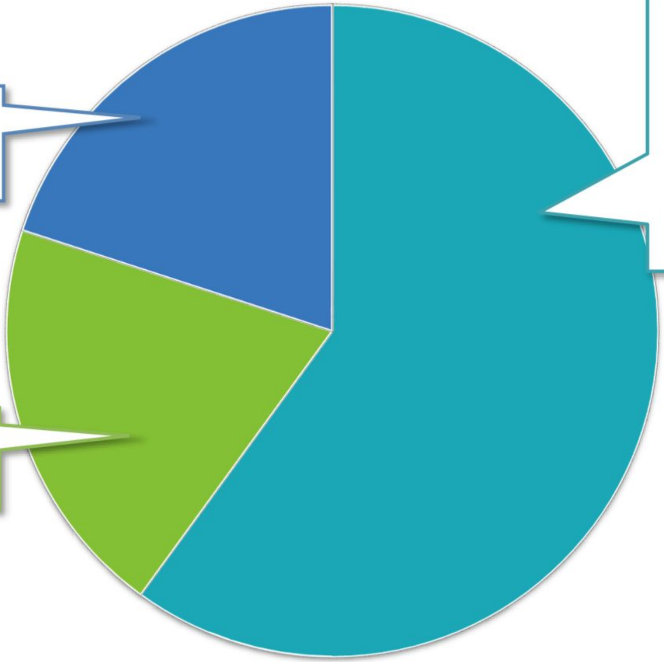
Slide courtesy Katherine Riley, ALCF

Argonne NATIONAL LABORATORY

Primary allocation programs for access to LCF in 2020
Current distribution of allocable hours

Slide courtesy Verónica G. Vergara Larrea, OLCF

20% Director's Discretionary
(Includes LCF strategic programs, ECP)



Up to 60% INCITE

Leadership Computing Class

20% ASCR Leadership Computing Challenge

DOE/SC capability computing

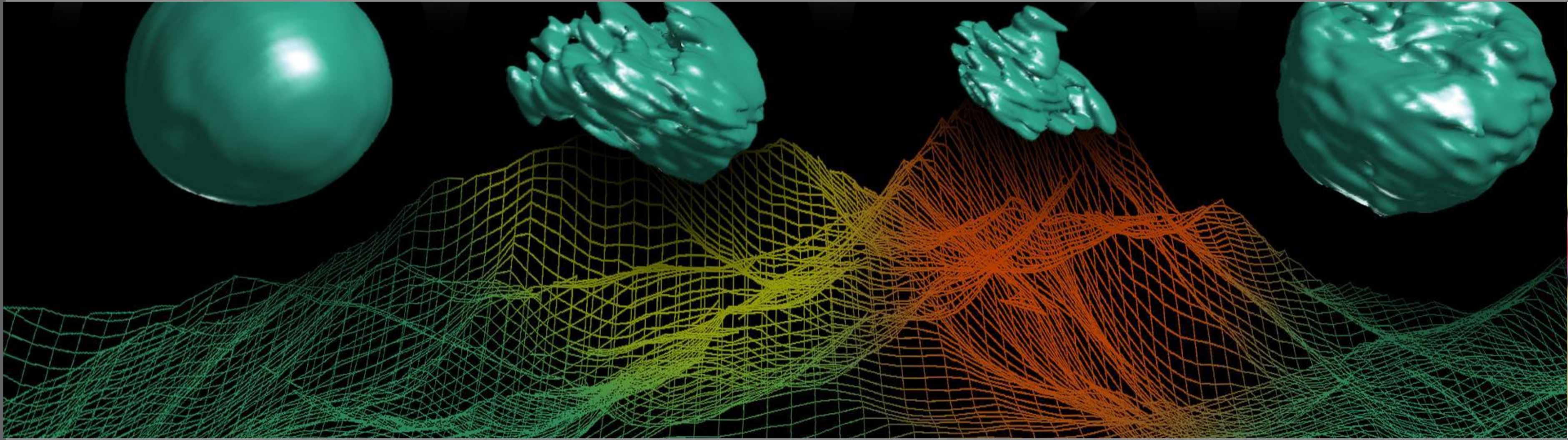
DD

- Smaller allocations (<5 Mch)
- Intended as onramp for new projects
- Preparation for larger allocation programs
- Proposals accepted year round (starting Dec. for Summit)

Getting Started (DD)

Our Director's Discretionary (DD) allocation program provides researchers with small awards of computing time to "get started" on our computing resources while pursuing real scientific goals.

The DD allocation program allows users to prep their code so that it can take advantage of our massively parallel systems.



Slide courtesy Katherine Riley, ALCF

DD
Director's Discretionary

Purpose: A "first step" for projects working toward a major allocation

Eligibility: Available to all researchers in academia, industry, and other research institutions

Review Process: Projects must demonstrate a need for high-performance computing resources; reviewed by ALCF

Award Size: Low 10 thousand of node-hours

Award Duration: 3-6 months, renewable

Total percent of ALCF resources allocated: 20%

Award Cycle

Ongoing (available year round)

Slide courtesy Katherine Riley, ALCF

Apply for allocations at the Leadership Computing Facilities

Oak Ridge Leadership Computing Facility

https://www.olcf.ornl.gov/for-users/getting-started/#request-a-new-allocation

Argonne Leadership Computing Facility

https://accounts.alcf.anl.gov/#!/allocationRequest

Americas HPC Collaboration

- BoF at SC19
  - Showcase collaboration opportunities and experiences between different HPC Networks and Laboratories from countries of the American continent
- Bof at SC20
  - TBA
- Join us !
  - https://join.slack.com/t/hpc-americas-collab/shared_invite/zt-g483zw52-JQIEf5NYtIwlqH5P6qA45Q

What’s Next?

Part 3

- Software Engineering - Best Practices
- Open-sourcing workflow
  - GitHub and Zenodo
- Publication process
  - Getting accepted in two journals
- Remote collaboration
  - How we work together without ever meeting in person, even though COVID-19
- Looking for highly motivated collaborators
- Q&A

Open Source is more than just uploading to GitHub

General

- README
- LICENSE
- Contributions/Contributors
- Conribution guidelines

Functionality

- Installation instructions
- Running tests or clear instructions for testing
- Performance

Other documentation

- Statement of need (the Why?)
- Example of how to use at basic level
- Community Guidelines

SE

- Automatic build script or Makefile
- Design/Architecture
- Unit Tests or Test Programs
- Regular Commits
- Pull Requests

We follow most of these on our project. These guidelines are based on JOSS journal, which is focused on peer-reviewed research software artifacts.

Zenodo

Zenodo is a service created by CERN for storing research artifacts (code, datasets, analysis).

CERN is where the WWW was created.

Original purpose of the web was to support the dissemination of scientific information.

Zenodo is basically bringing the web back to first principles.

Anyone with a valid account and scientific/research purpose can upload datasets up to 50GB

We use Zenodo to store our data sets and analysis results and all software

Why use Zenodo? (see zenodo.org for details)

- Safe — your research is stored safely for the future in CERN’s Data Centre for as long as CERN exists.
- Trusted — built and operated by CERN and OpenAIRE to ensure that everyone can join in Open Science.
- **Citeable — every upload is assigned a Digital Object Identifier (DOI), to make them citable and trackable.**
- No waiting time — Uploads are made available online ...
- Open or closed — Share e.g. anonymized clinical trial data with only medical professionals via our restricted access mode.
- **Versioning — Easily update your dataset with our versioning feature.**
- **GitHub integration — Easily preserve your GitHub repository in Zenodo.**
- Usage statisics — All uploads display standards compliant usage statistics

Zenodo artifact from GitHub Release(s)

Dematties, Dario, Thiruvathukal, George K., Rizzi, Silvio, Perez, Mauricio D., Wainselboim, Alejandro Javier, & Zanutto, Bonifacio Silvano. (2019, August 22). neurophon/neurophon: A Computational Theory for the Emergence of Grammatical Categories in Cortical Dynamics (Version v1.2). Zenodo. http://doi.org/10.5281/zenodo.3374889

- GitHub Integration for Zenodo
- Author metadata taken from .zenodo.json (top level of the repo, next slide)
- Each version of the software gets a unique DOI
- git tag <version> automatically pushes to Zenodo with this integration.

```json
{
  "creators": [
    {
      "name": "Dematties, Dario",
      "affiliation": "University of Buenos Aires",
      "orcid": "0000-0002-8726-7837"
    },
    {
      "name": "Thiruvathukal, George K.",
      "affiliation": "Loyola University Chicago and Argonne National Laboratory",
      "orcid": "0000-0002-0452-5571"
    },
    {
      "name": "Rizzi, Silvio",
      "affiliation": "Argonne National Laboratory",
      "orcid": "0000-0002-3804-2471"
    },
    // List of authors shortened for presentation purposes.
  ],
  "keywords": [
    "cortical dynamics",
    "early language acquisition",
    "incidental phonetic acquisition",
    "sparse distributed representations",
    "unsupervised learning",
    "biologically inspired computational models",
    "neural networks"
  ],
  "license": "GPL-3.0",
  "upload_type": "software"
}
```

.zenodo.json

Zenodo examples for storing datasets/analysis

Dario Dematties, Silvio Rizzi, George K. Thiruvathukal, Alejandro Javier Wainselboim, Bonifacio Silvano Zanutto, & Mauricio D. Perez. (2019). A Computational Theory for the Emergence of Grammatical Categories in Cortical Dynamics [Data set]. Zenodo. http://doi.org/10.5281/zenodo.3653180

Dematties, Dario, Thiruvathukal, George K., Rizzi, Silvio, Wainselboim, Alejandro Javier, & Zanutto, Bonifacio Silvano. (2019). Experimental Results and Appendices: Cortical Spectro-Temporal Model (CSTM). [Data set]. Zenodo. http://doi.org/10.5281/zenodo.2654939

- These are manually created on Zenodo (not GitHub based)
- Most research software makes use of larger files best kept outside of git repo
- Zenodo lets you store 50GB of data. We reference the Zenodo archives in our journal submissions and in GitHub docs so othres can reproduce our study.

PLOS ONE

RESEARCH ARTICLE

Phonetic acquisition in cortical dynamics, a computational approach

Dario Dematties[1]*, **Silvio Rizzi**[2], **George K. Thiruvathukal**[2,3], **Alejandro Wainselboim**[5], **B. Silvano Zanutto**[1,4]

1 Universidad de Buenos Aires, Facultad de Ingeniería, Instituto de Ingeniería Biomédica, Ciudad Autónoma de Buenos Aires, Argentina, **2** Argonne National Laboratory, Lemont, Illinois, United States of America, **3** Computer Science Department, Loyola University Chicago, Chicago, Illinois, United States of America, **4** Instituto de Biología y Medicina Experimental-CONICET, Ciudad Autónoma de Buenos Aires, Argentina, **5** Instituto de Ciencias Humanas, Sociales y Ambientales, Centro Científico Tecnológico-CONICET, Ciudad de Mendoza, Mendoza, Argentina

* ddematties@fi.uba.ar

Check for updates

Abstract

frontiers in Neural Circuits

ORIGINAL RESEARCH
published: 16 April 2020
doi: 10.3389/fncir.2020.00012

Check for updates

A Computational Theory for the Emergence of Grammatical Categories in Cortical Dynamics

Dario Dematties[1*], *Silvio Rizzi*[2], *George K. Thiruvathukal*[2,3], *Mauricio David Pérez*[4], *Alejandro Wainselboim*[5] *and B. Silvano Zanutto*[1,6]

[1] Universidad de Buenos Aires, Facultad de Ingeniería, Instituto de Ingeniería Biomédica, Buenos Aires, Argentina, [2] Argonne National Laboratory, Lemont, IL, United States, [3] Computer Science Department, Loyola University Chicago, Chicago, IL, United States, [4] Microwaves in Medical Engineering Group, Division of Solid-State Electronics, Department of Electrical Engineering, Uppsala University, Uppsala, Sweden, [5] Centro Científico Tecnológico Conicet Mendoza, Instituto de Ciencias Humanas, Sociales y Ambientales, Mendoza, Argentina, [6] Instituto de Biología y Medicina Experimental-CONICET, Buenos Aires, Argentina

Parallel Computing: Technology Trends
I. Foster et al. (Eds.)

doi:10.3233/APC200077

Towards High-End Scalability on Biologically-Inspired Computational Models

Dario DEMATTIES [a], George K. THIRUVATHUKAL [b,c] Silvio RIZZI [c]
Alejandro WAINSELBOIM [e] B. Silvano ZANUTTO [a,d]

[a] *Universidad de Buenos Aires, Facultad de Ingeniería, Instituto de Ingeniería Biomédica, Ciudad Autónoma de Buenos Aires, Argentina*
[b] *Computer Science Department, Loyola University Chicago, Chicago, Illinois, United States*
[c] *Argonne National Laboratory, Lemont, Illinois, United States*
[d] *Instituto de Biología y Medicina Experimental-CONICET, Ciudad Autónoma de Buenos Aires, Argentina*
[e] *Instituto de Ciencias Humanas, Sociales y Ambientales, Centro Científico Tecnológico-CONICET, Ciudad de Mendoza, Mendoza, Argentina*

¡Muchas Gracias!