OcéanIA
AI, Oceans and Climate Change

Nayat Sánchez Pi
Director Inria Research Center in Chile
Diversity in Natural Resources: Atacama desert
Challenging enough to AI/ML

- Small, big and/or high dimensional data
- Transfer learning
- Causal Learning
- Active Learning
- Explainable AI

Interdisciplinary

ASTRONOMY

ENVIRONMENT

➔ Inherent complex context
➔ Convergence of physical and life sciences through Big and Data
Inria Challenge

OcéanIA: Artificial Intelligence and Modelling for Understanding Climate Change

https://oceania.inria.cl/
Nature has provided a lot of inspiration for the AI/ML area:

- neural networks,
- learning theories,
- evolutionary computing, etc.

It is about time that we return the favor!
Climate Change and the Ocean

Environment

- **TARA ARCTIC (2006-2008)**: First Arctic drift since Nansen (1893)
- **TARA OCEANS (2009-2013)**: First global study of the planktonic ecosystem
- **TARA MEDITERRANEAN (2014)**: Study of the impact of plastic on the marine ecosystem in the Mediterranean
- **TARA PACIFIC (2016-2018)**: Study of the adaptive capacity of coral reefs to climate change
- **MICROPLASTICS (2019)**: First study of river sources of microplastics on a European scale

11 missions including 5 major expeditions
100,000 microscopic marine species discovered
Over 150 million marine genes discovered
Almost 200,000 viruses characterized
Over 150,000 samples collected
Expedition
Feb - May 2021

Mission Microbiome - CEODOS

Understand the ocean environment, climate change, its impact and mitigation.
Why Tara?

Schooner Tara:

- >40,000 plankton samples collected
- 210 sampling stations at three depths
- >60 terabases of DNA and RNA sequenced
- ~7 million images captured

⇒ Tara Oceans adopts the principle of open access and early release of raw and validated data sets.

⇒ Solid, diverse and international scientific community.

OcéanIA

An opportunity and a challenge to state-of-the-art AI/ML
Small but high-dimensional heterogeneous data

- Tara expeditions gather lots of data - from a marine biology point of view.
- Highly heterogeneous: DNA barcodes, images, environment variables.
- Samples includes many species at the same time.
- Grouped by sample location, interested in networks and graph-based information.
- In spite of efforts it is not always consistent and it is always evolving: i.e. new hardware.
- Data from Tara allows exploration of the relationship between marine ecosystem functioning and biodiversity.

Model reuse, transfer learning and domain adaptation

A three-step work hypothesis:

↳ Is it possible to adapt existing (computer vision or graph) models to out domain?
↳ How to transfer a domain model to other biomes, locations or across species?
↳ How to cope with variations across species and sensing hardware?

● **Transfer learning** addresses the issue of how to adapt and re-purpose the internal representations of a model that has been trained on a similar problem.

● **Domain adaptation** is the capacity to cope with changes in the environment because of the evolution of the system and/or the need to particularize a general model to a particular instance.
Active, few-shot and multi-task learning

- Limited data and/or high uncertainty,
- Direct sampling to the areas of the domain where they are most necessary.
- Guiding sampling using **active learning**.
- **Few-shot learning** methods to produce actionable products with minimal data.
- **multi-source** or **multi-task learning** ensembles training signals of related tasks.
  - Enables the model to generalize better on the main task.
  - Effectively increases the sample size that is being used for training.
  - Biases the model to prefer representations that are useful for other tasks -> transfer learning!

Source: https://brandidea.ai/activeLearning.html
Causality and Explainable AI

→ Produce explainable models, while maintaining performance (prediction accuracy),

→ research support tools that combine explainability and causality for new scientific discoveries and theories by making surrogate human-readable models, and

→ enable human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners.

Essential for policy making.

Biodiversity supports functions like primary productivity and carbon fixation and sequestration, etc.

Understanding this is fundamental: science and policy making.

Data from Tara allows exploration of the relationship between marine ecosystem functioning and biodiversity.

How variations on biodiversity impact those functions?

How changes in temperature (or other variables) impact biodiversity and functions?

Understand causality and circular causality among different levels of biodiversity, ecosystem functioning.

Understanding plankton communities using AI, ML, and vision

Mapping images to function (A→D)

Mapping images to individuals (A→I)
Anomaly detection and explainable AI for automatic plankton ID

→ Identifying plankton as a supervised problem: already addressed.

→ Estimated that more than 70,000 species unknown. ⇒ unsupervised or active learning approaches.

→ Tara sampling includes high-res microscope.

→ Identify unknown or out of context species.

→ Why an organism represents an interesting specimen?

→ Methods involved:
  ⇒ transfer learning and domain adaptation,
  ⇒ (un/self)supervised object detection and segmentation
  ⇒ causal inference to understand context.
  ⇒ explainable AI: i.e. hint what parts of the organism that determining its selection.

Field experts want a research tool not a black-box.
Integrating model-based (i.e. PDEs) and data-based (i.e. ML) approaches

Ocean Circulation Models studying dynamics:

- high-res models are computationally inviable,
- current resolution of models is not sufficient,
- require large viscosity and diffusion coefficients smooths out features such as jets and mesoscale eddies.
- oceanic turbulence at small scales, which play an important role.

Planned actions:

- Learning PDEs from Data
- Understanding learning dynamics
- Hybrid models: Combining PDE solvers and DNNs (improved explicability).
Structure and graph-based NNs

Capacity of coupling complex and structured information with powerful the machine learning method

Learning and adaptation

Models learn simultaneously different tasks, a major challenge and with several applications

Causal inference and explainable AI

Understanding nature and at the same time to be a source for new theories

Model-/data-driven integration and hybrids

Repurpose models with minimal impact, i.e., study new species, other regions, etc.

Development, calibration and validation of mechanistic models

Problems with limited data and/or high uncertainty need direct the sampling where it is most needed

Integrating biodiversity community structures and function along the ocean

Computer vision for understanding plankton communities
The OcéanIA team

https://oceania.inria.cl/

Data Engineering and HPC
Data Governance
Artificial Intelligence
Causal Learning
PDEs
Mathematical Modelling

Inria Challenge

LEADER
France
Chile

ANGE
TAU
BIOCORE
GO-SEE
Federation

Others

Support

COVID-19

NEURAL INFORMATION PROCESSING SYSTEMS
LXAI

December 7th, 2020
Oceans and climate change are intimately related:
  > Carbon capture, impact of change of temperatures, etc.

Oceans are the last 'unknown':
  > Understanding the role of oceans in climate change is not only important but also challenging for modern AI, ML and applied math.

A way to address current hot topics like explainability, bias, etc. on a different domain.

Tara as source of data that can be crossed with other sources.

I have described the approaches we are to follow in the next 4 years.

OcéanIA has just started!
  > We are hiring!
Thank you! Merci !
Obrigada! ¡Gracias!

Follow us: www.inria.cl and oceania.inria.cl