Data-centric Machine Learning Applied to Sustainability

Laure Berti-Equille

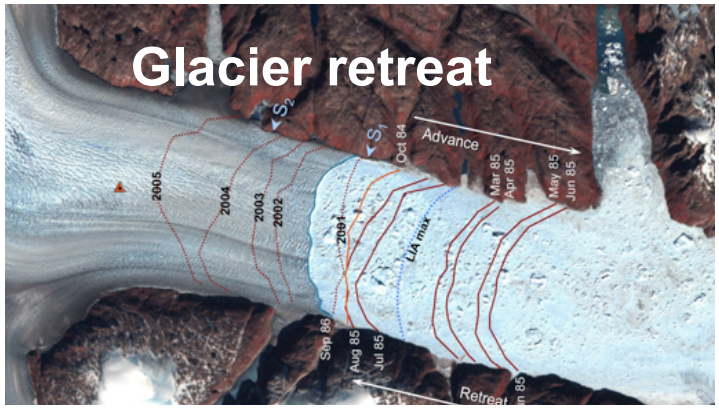
IRD ESPACE-DEV
Montpellier, France
laure.berti@ird.fr
https://laureberti.github.io/website/

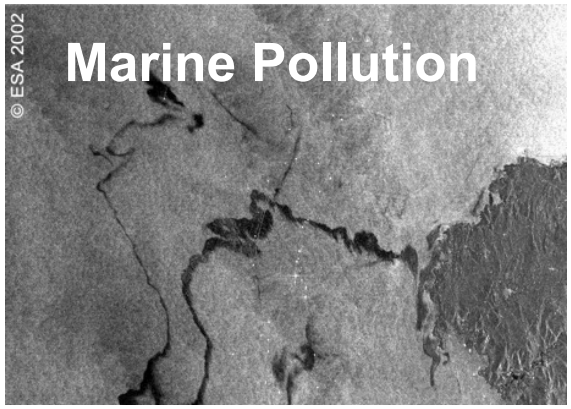
Institut de Recherche pour le Développement
FRANCE
https://en.ird.fr/

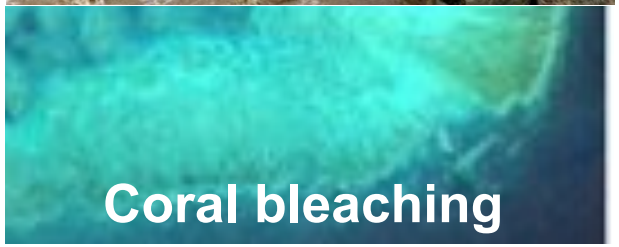
IDEAL
Agroecology & AI Joint Lab

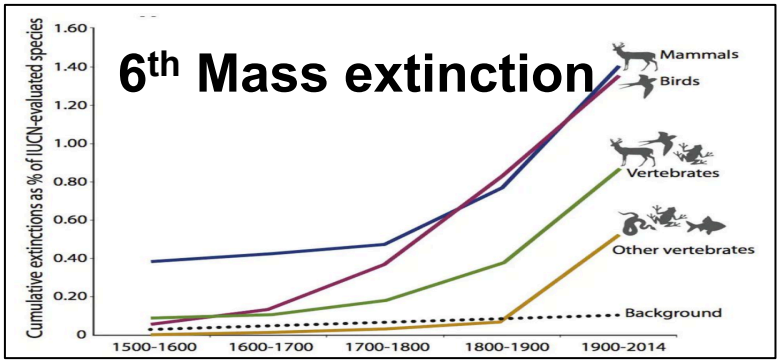
Espace DEV
OBSERVATION SPATIALE, MODÈLES & SCIENCE IMPLIQUÉE
https://www.espace-dev.fr/en/

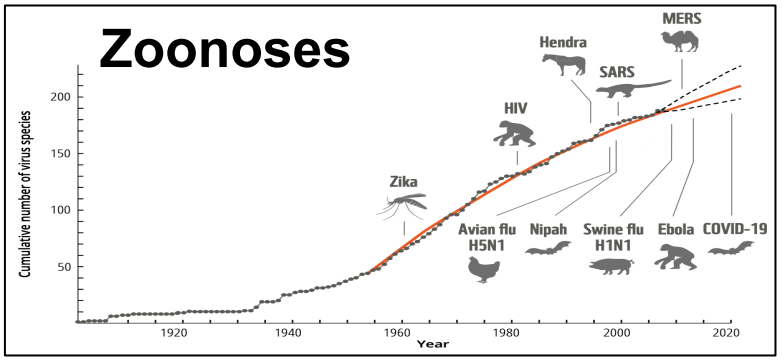
Today's View of Our World











2030 Sustainable Development Goals

2030 Agenda for Sustainable Development: 17 goals, 169 targets, 232 Indicators

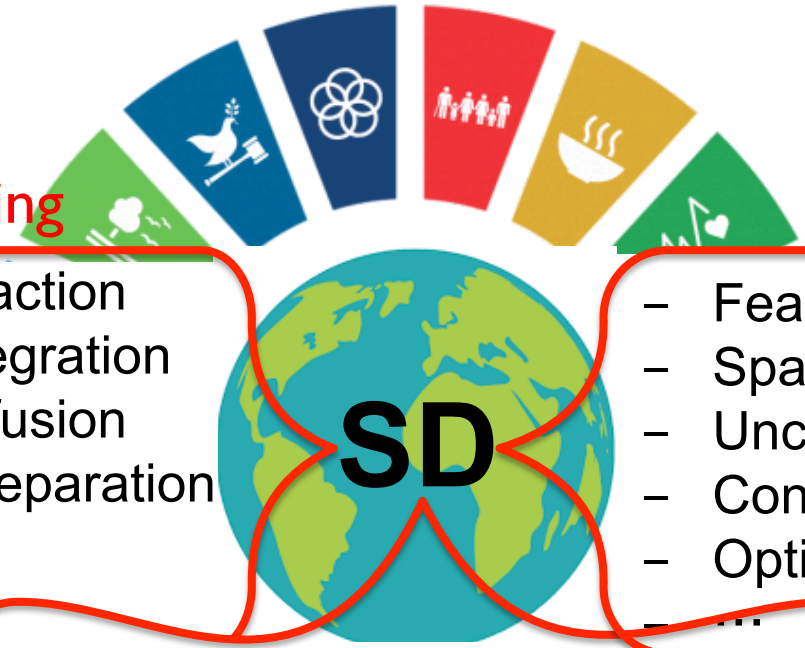
New norms to integrate the principles of sustainable development into country policies and programs

https://sdgs.un.org/2030agenda

https://sdgs.un.org/goals

Applying Data Science & AI to SDGs

Transdisciplinary research, reasoning, and discovery from interconnected information contents



Data Engineering

- Information extraction
- Data linking, integration
- Multiscale data fusion
- Data curation/preparation
- Data analytics
- …

Model Engineering

- Feature engineering
- Spatiotemporal & dynamic models
- Uncertainty quantification
- Constraint reasoning
- Optimization
- …

Citizen Science

- Crowd sensing/sourcing
- Behavioral game theory
- HITL
- …

A wide range of non-CS disciplines: Earth, Environmental, Social sciences, etc.

A Decade of AI Contributions & Initiatives

Tackling Climate Change with Machine Learning

David Rolnick[1]*, Priya L. Donti[2], Lynn H. Kaack[3], Kelly Kochanski[4], Alexandre Lacoste[5], Kris Sankaran[6,7], Andrew Slavin Ross[9], Nikola Milojevic-Dupont[10,11], Natasha Jaques[12], Anna Waldman-Brown[12], Alexandra Luccioni[6,7], Tegan Maharaj[6,8], Evan D. Sherwin[2], S. Karthik Mukkavilli[6,7], Konrad P. Körding[1], Carla Gomes[13], Andrew Y. Ng[14], Demis Hassabis[15], John C. Platt[16], Felix Creutzig[10,11], Jennifer Chayes[17], Yoshua Bengio[6,7]

nature communications

Explore content · Journal information · Publish with us

nature > nature communications > perspectives > article

Perspective | Open Access | Published: 13 January 2020

The role of artificial intelligence in achieving the Sustainable Development Goals

Ricardo Vinuesa, Hossein Azizpour, Iolanda Leite, Madeline Balaam, Virginia Dignum, Sa
Domisch, Anna Felländer, Simone Daniela Langhans, Max Tegmark & Francesco Fuso Ner

CompSustNet
Computational Sustainability Network

www.compsust.net

Climate Change AI

www.climatechange.ai

SDGacademy
https://sdgacademy.org/

Climate Informatics

www.climateinformatics.org

KDD 2017 Applied Invited Talk — KDD'17, August 13–17, 2017, Halifax, NS, Canada

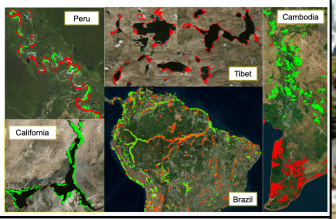
Big Data in Climate: Opportunities and Challenges for Machine Learning

Anuj Karpatne
Department of Computer Science and Engineering, University of Minnesota
karpa009@umn.edu

Vipin Kumar
Department of Computer Science and Engineering, University of Minnesota
kumar001@umn.edu

ABSTRACT

The climate and Earth sciences have recently undergone a rapid transformation from a data-poor to a data-rich environment. In particular, massive amount of data about Earth and its environment is now continuously being generated by a large number of Earth observing satellites as well as physics-based earth system models running on large-scale computational platforms. These massive and information-rich datasets offer huge potential for understanding how the Earth's climate and ecosystem have been changing and how they are being impacted by humans actions. We discuss the challenges involved in analyzing these massive data sets as well as opportunities they present for both advancing machine learning as well as the science of climate change.



ENERGY

How A.I. Will Revolutionize Climate Tech

This week on The Interchange: we look deeper at artificial intelligence as a climate

STEPHEN LACEY | JUNE 17, 2021

APPLIED ECONOMIC PERSPECTIVES AND POLICY
AAEA Agricultural & Applied Economics Association

Submitted Article

Big Data in Agriculture: A Challenge for the Future

Keith H Coble, Ashok K Mishra, Shannon Ferrell, Terry Griffin

First published: 16 February 2018 | https://doi.org/10.1093/aepp/ppx056 | Citations: 10

4

5

SD as Optimization Problems

- Maximizing the probability of achieving an SD target
- Minimizing the degradations of environmental and human conditions

Complexity:

- Multi-objective: improve the quality of human life, preserve the Earth's diversity, minimize the depletion of non-renewable resources
- Multi-disciplinary: environmental, social sciences, etc.
- Multi-scale in time and space: global, national, regional, local with various horizons
- Multi-actor: civil society, private companies, government

ML for SD: Main Challenges

Data

1. Lack of representative ground truth data
2. Dirty and imperfect data with data quality issues, uncertainty, label noise, bias

Models

3. Inappropriate feature engineering and inadequate data preprocessing
4. Energy efficiency and carbon footprint reduction of ML models

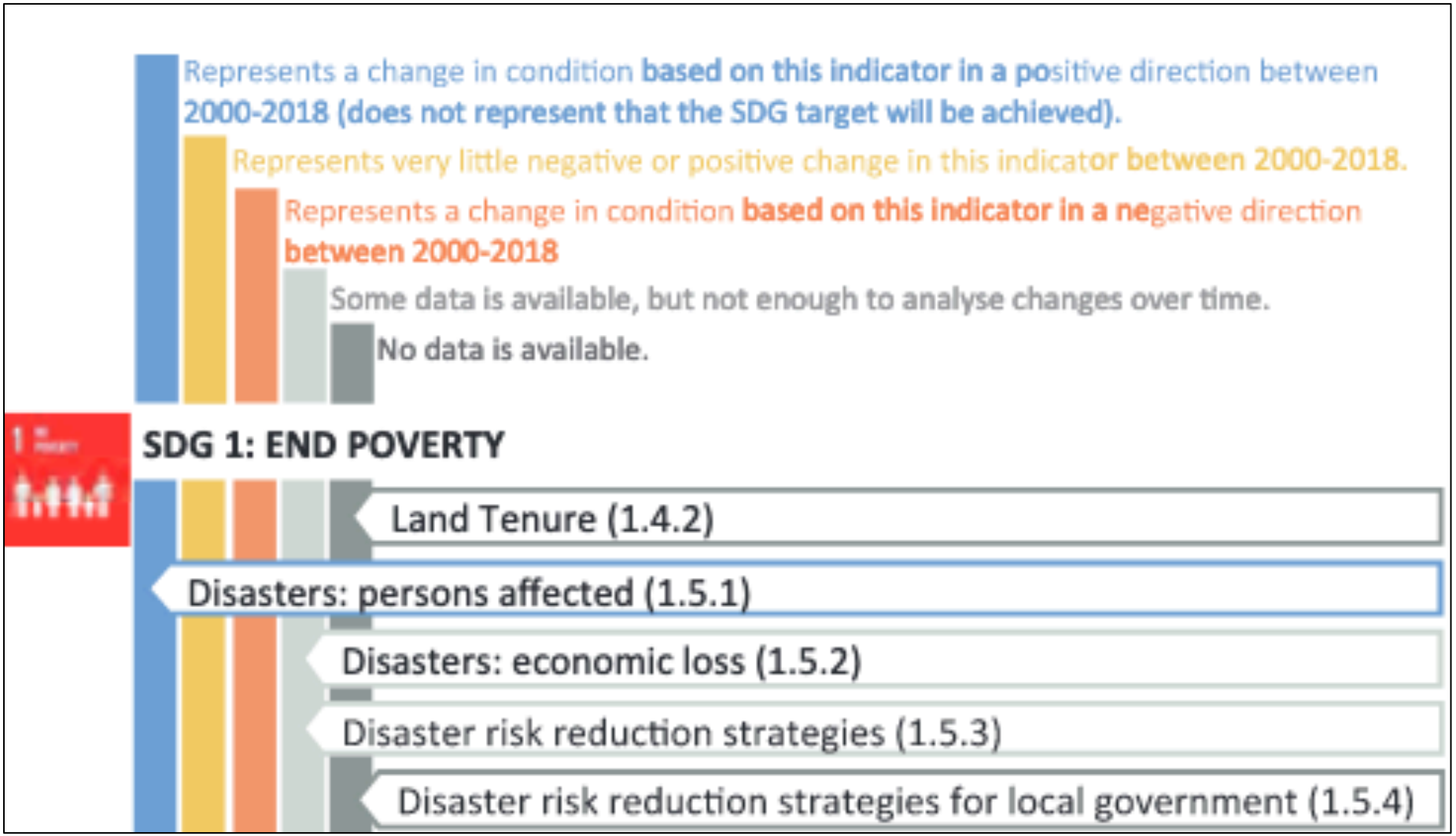
Output

5. Non actionable and misleading outputs
6. Difficult to measure the impact of ML-based solutions
7. Interpretability and trust issues for local actors

Progress on the SDGs

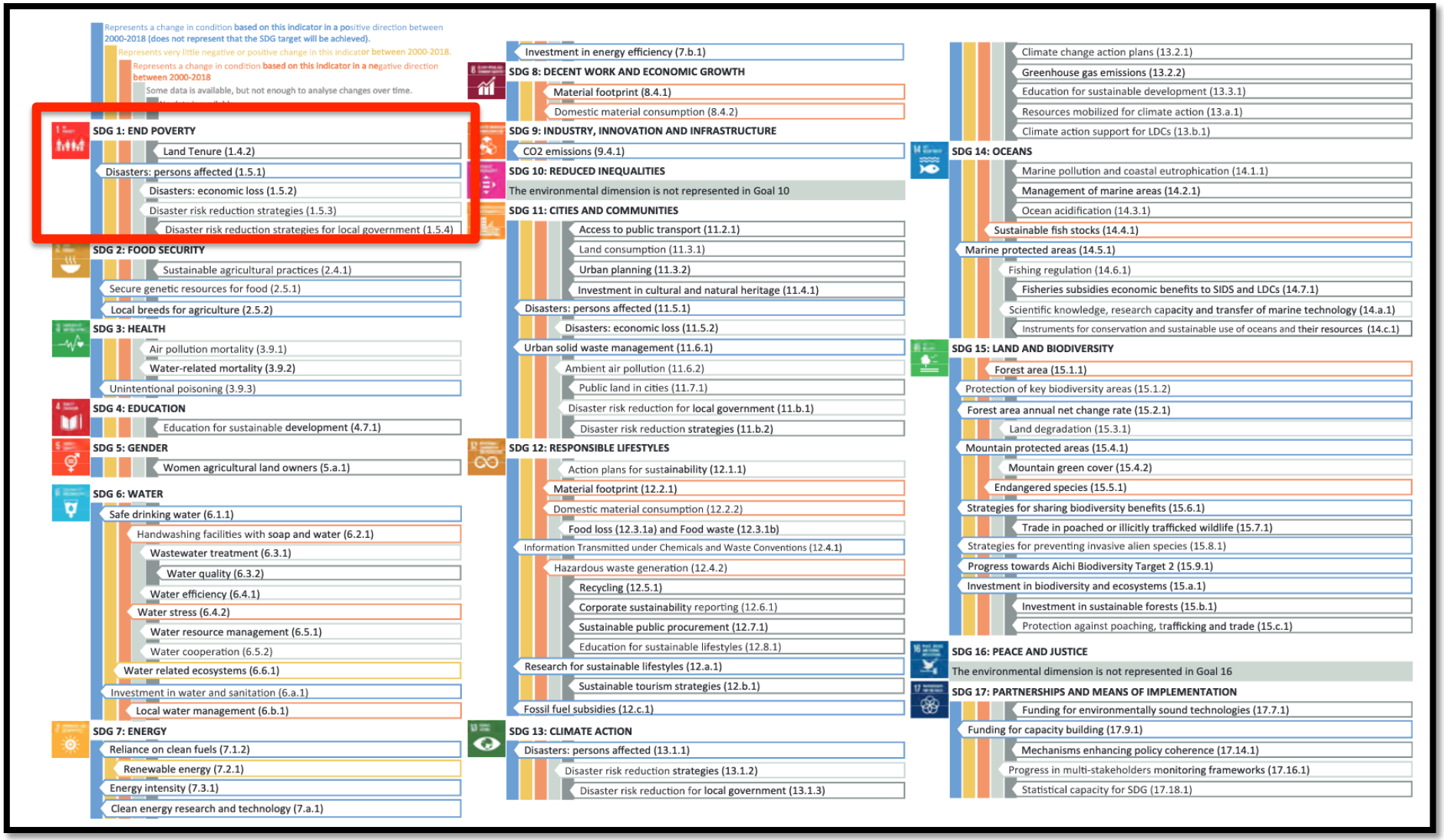
https://sdg-tracker.org/

Example: The Lack of Data



https://www.unep.org/resources/publication/measuring-progress-environment-and-sdgs

Progress on the SDGs



https://www.unep.org/resources/publication/measuring-progress-environment-and-sdgs

Outline

I- Challenges in ML & data science pipelines

- Building the pipelines
- Preparing the data

II- Current Projects

- Combining satellite imagery and socio-economic indicators to estimate poverty evolution in Africa
- Finding sustainable transition pathways in Nordeste

Illustrative Example 1

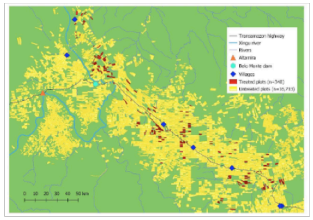
1. SD Question
2. Actors
3. Time/space of Interest
4. Data & Knowledge

Multi-Scale

Brazil

Minas Gerais

Area of study



Multi-Actor

at country-scale

- National authorities
- Policymakers
- Scientific community
- General public

at state-scale

- Federal authorities
- Regional actors

at territory-scale

- Administrator representatives of the population (stakeholders private and public)
- Site managers
- Local administrations
- Elected representatives of the local authorities
- Agroforestry exploitation owners
- Farmers
- Indigenous representatives

Illustrative Example 1

1. SD Question

   What are the fines effective for reducing deforestation?

2. Actors

   Coffee farmers, policy makers, local administrators

3. Time/space of Interest

   2018-2022, Minas Gerais, Brazil

4. Data & Knowledge

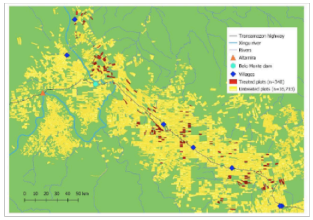
   Satellite images, surveys, agroecology datasets, etc.

Multi-Scale

Brazil

Minas Gerais

Area of study



Multi-Actor

at country-scale

- National authorities
- Policymakers
- Scientific community
- General public

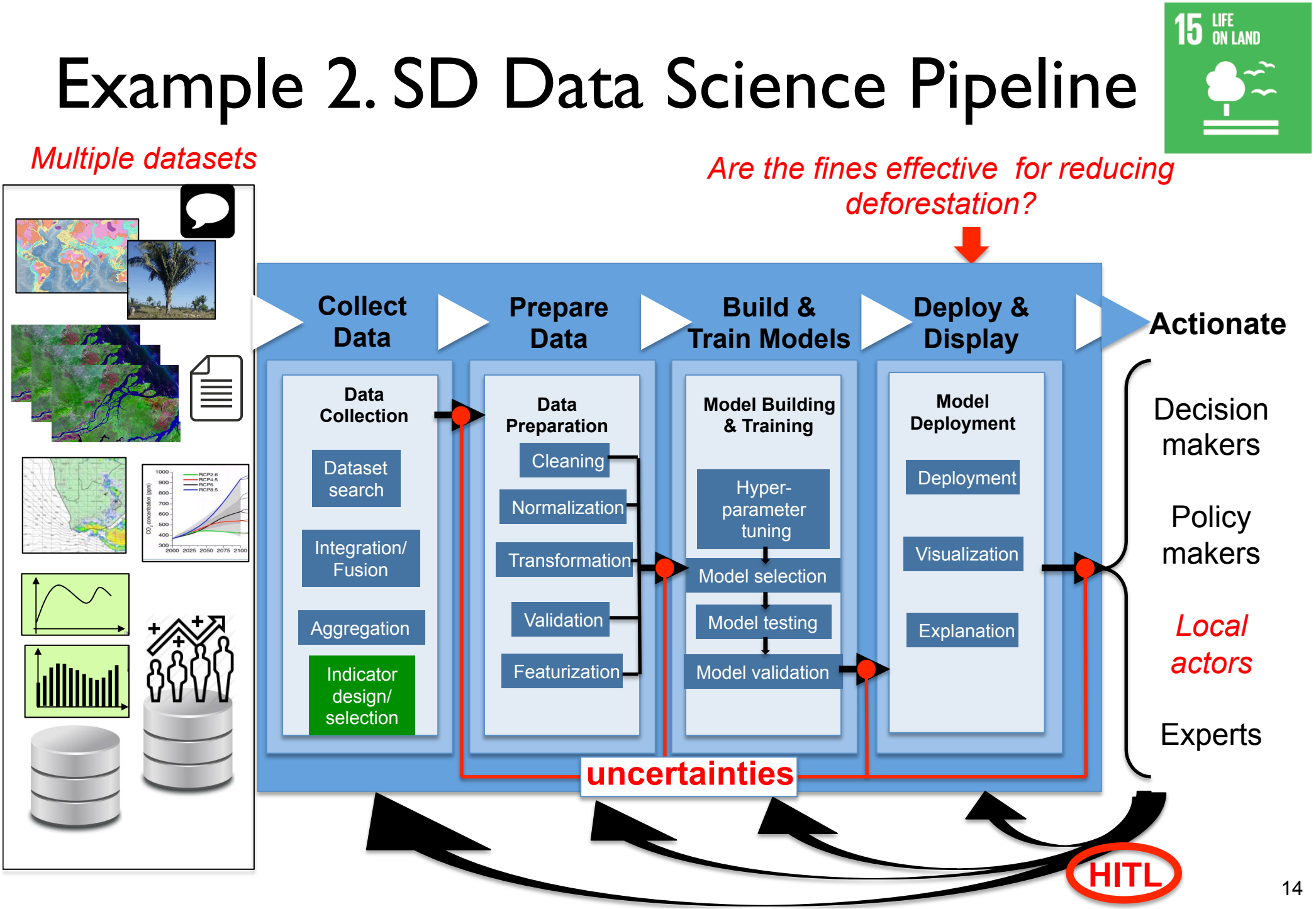
at state-scale

- Federal authorities
- Regional actors

at territory-scale

- Administrator representatives of the population (stakeholders private and public)
- Site managers
- Local administrations
- Elected representatives of the local authorities
- Agroforestry exploitation owners
- Farmers
- Indigenous representatives

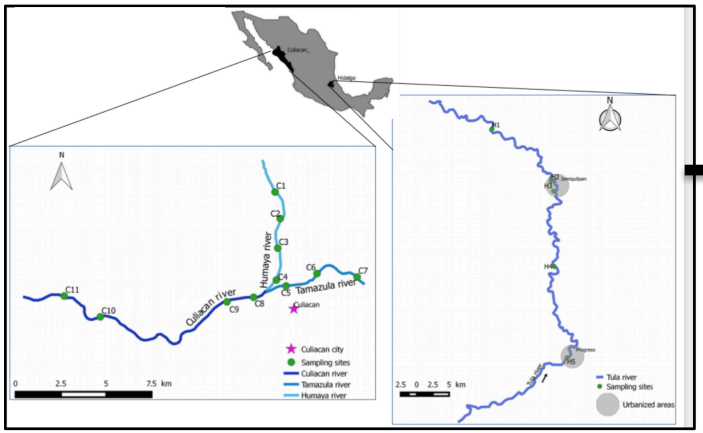
Example 2. SD Data Science Pipeline

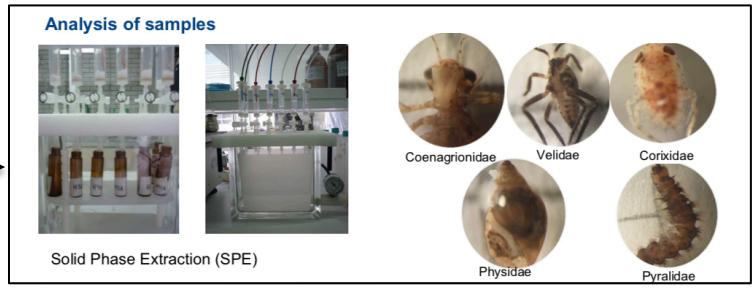


Data prep is crucial (1/2)

Are the Mexican rivers of Humaya, Tula, Tamazula, and Culiacan polluted ?

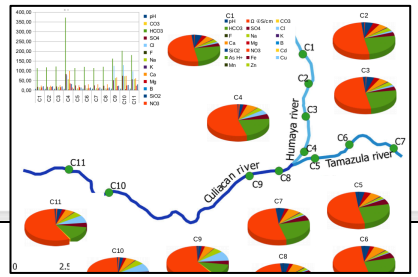
[Serrano Balderas et al. 2015, 2017]

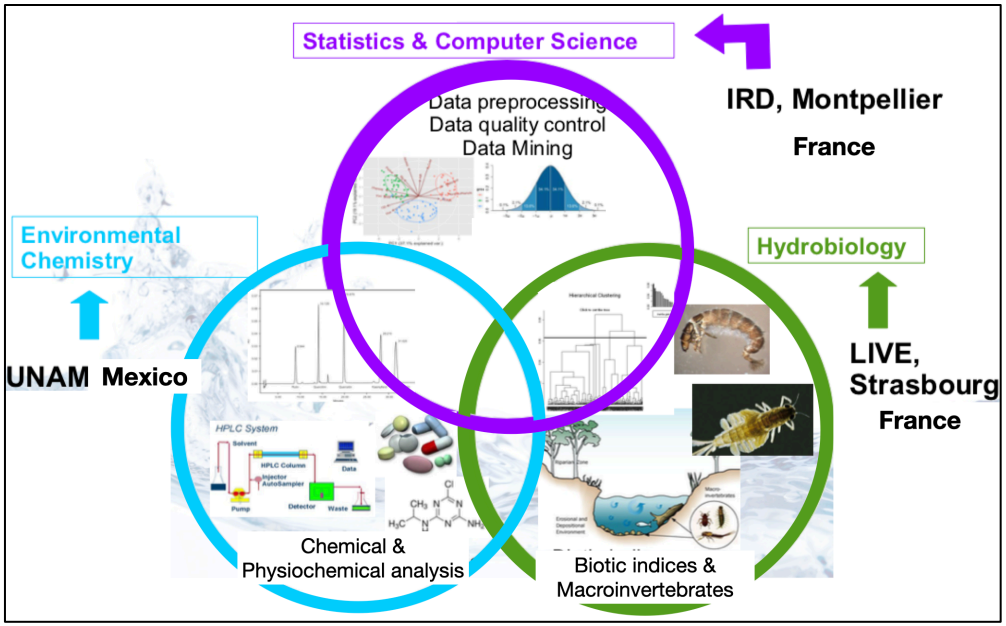




Water quality assessment using:

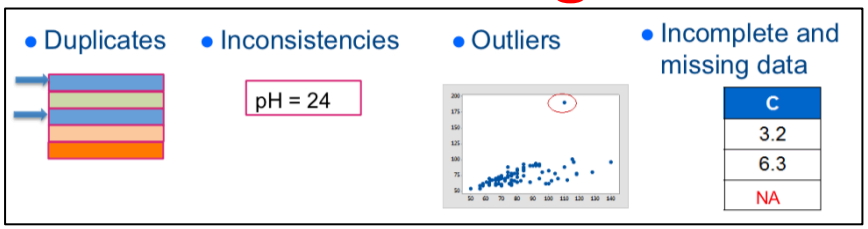
- **43** Physical-chemical variables
  - pH
  - Metals
  - Nitrates
  - Phosphates
  - Pesticides, etc.
- **35** Biological indicators based on the presence/ absence of aquatic macroinvertebrates
  - Measures of richness and enumerations
  - Diversity and similarity indices
  - Biotic indices
  - Functional feeding groups measures
  - Multimetric approach



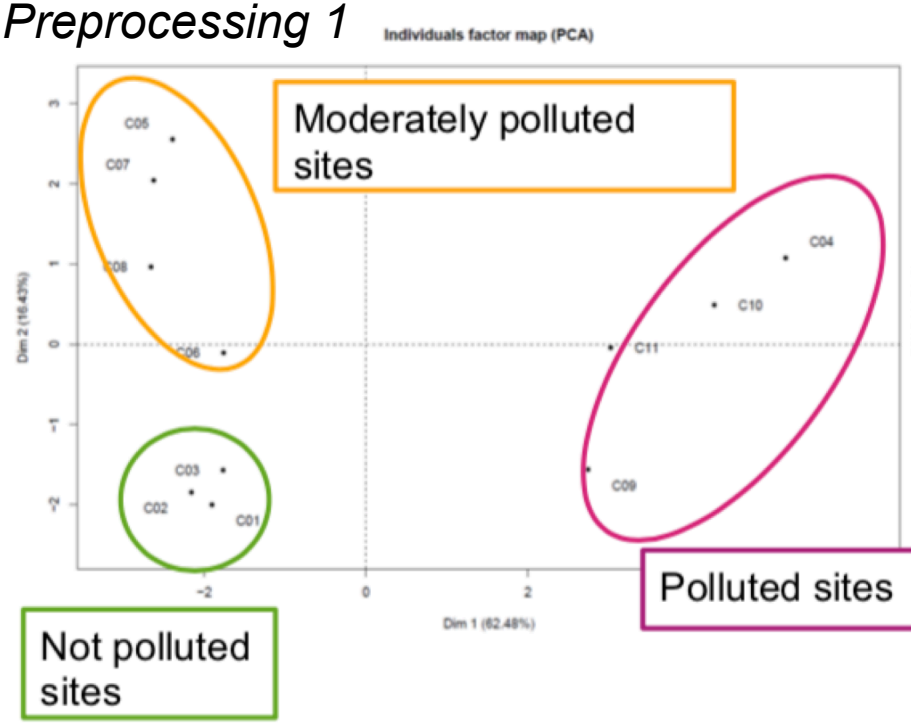


Data prep is crucial (2/2)

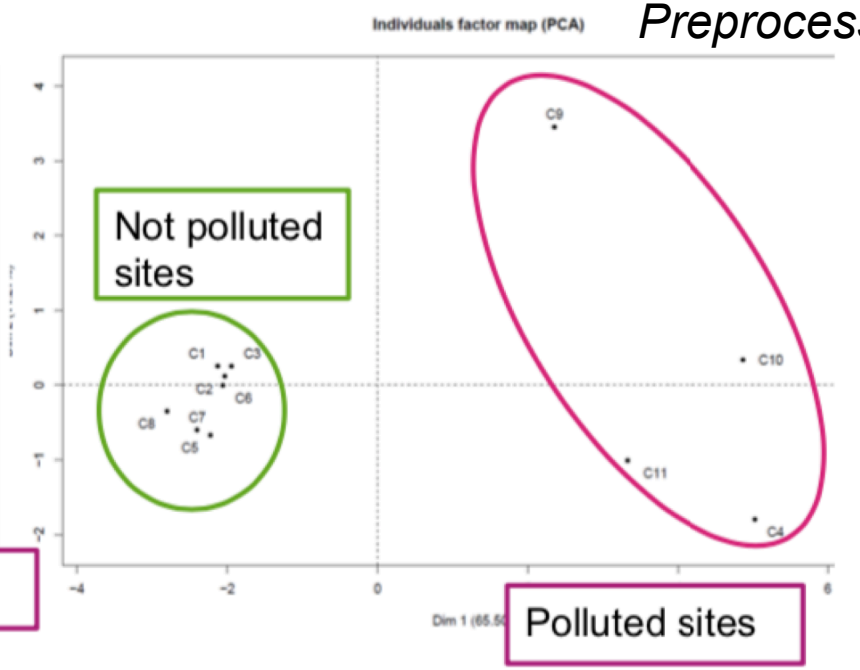
Various data cleaning/preparation strategies lead to different and misleading conclusions



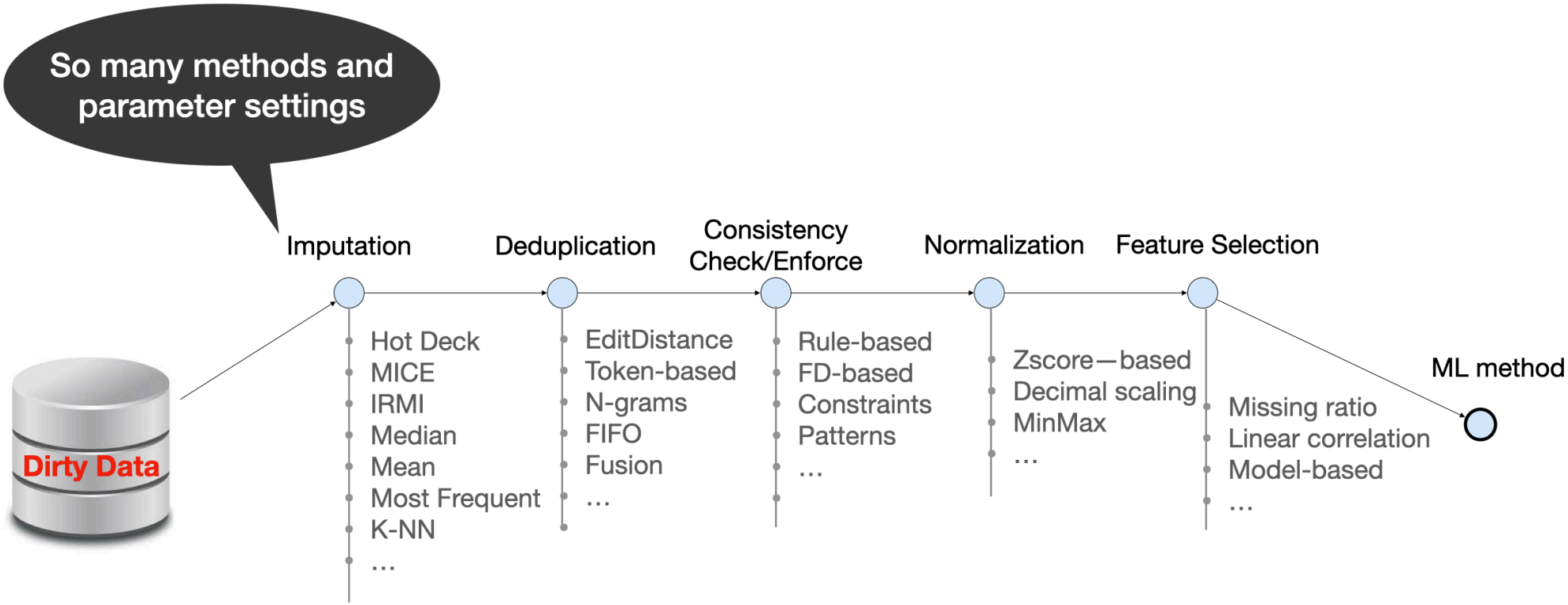
Preprocessing 1



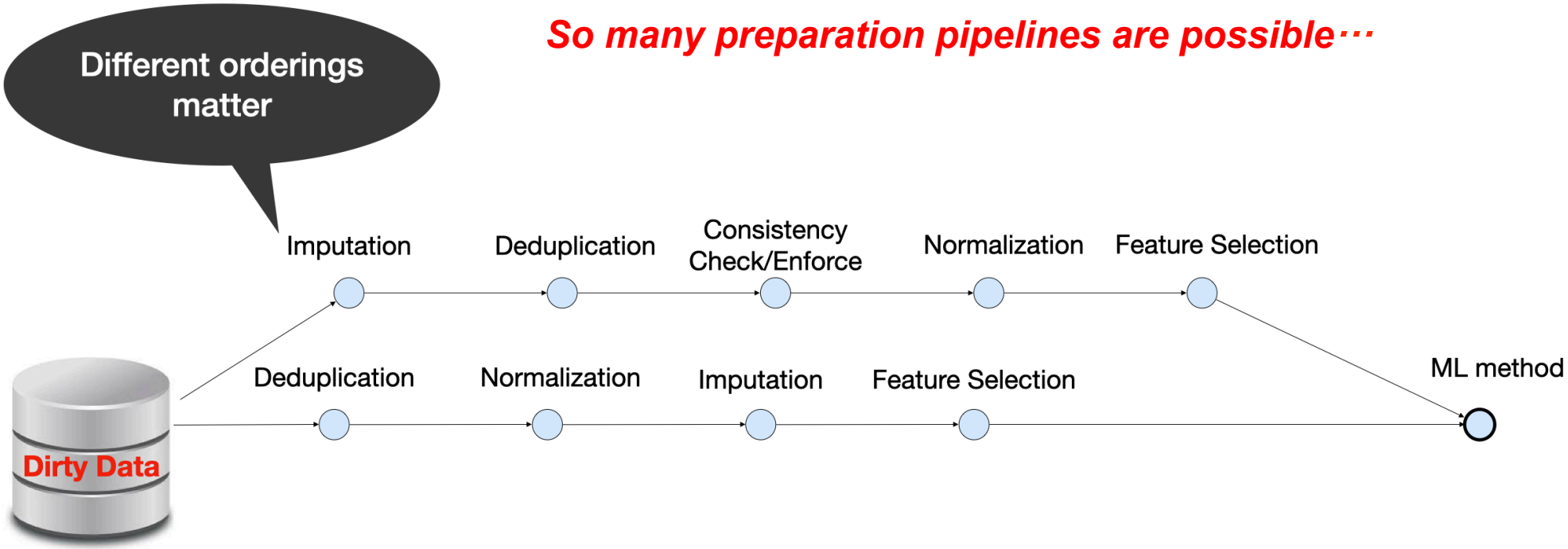
Preprocessing 2



Data prep is challenging



Data prep is challenging

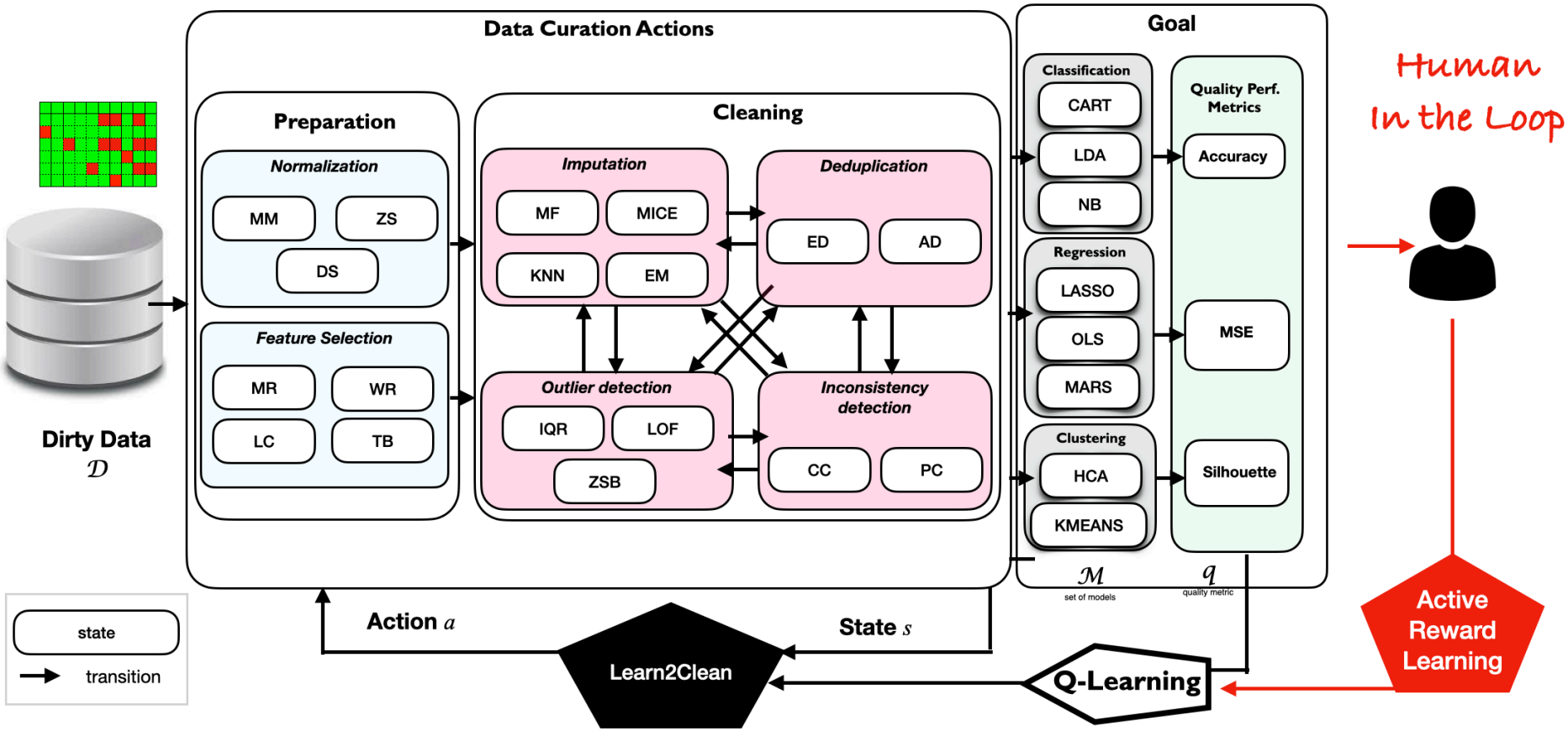


Reiterate some steps is possible…

Selective prep is also possible
… and even recommended

Reinforcement Learning can help data prep

[Berti-Equille, 2019, 2020]



https://github.com/LaureBerti/Learn2Clean

Outline

I- Challenges in ML & data science pipelines

- Building the pipelines
- Preparing the data

II- Current Projects

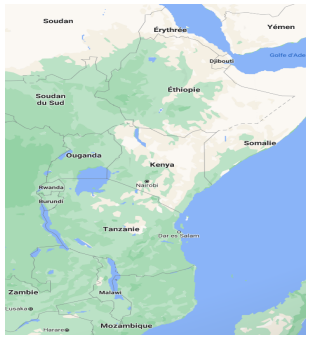
- Combining satellite imagery and socio-economic indicators to estimate poverty evolution in Africa
- Finding sustainable transition pathways in Brazil

Predict Poverty Evolution from Series of Satellite Images

MPA Poverty
funded by ANR

- **Incomplete surveys from 2009 to 2016 for 4 African Countries**

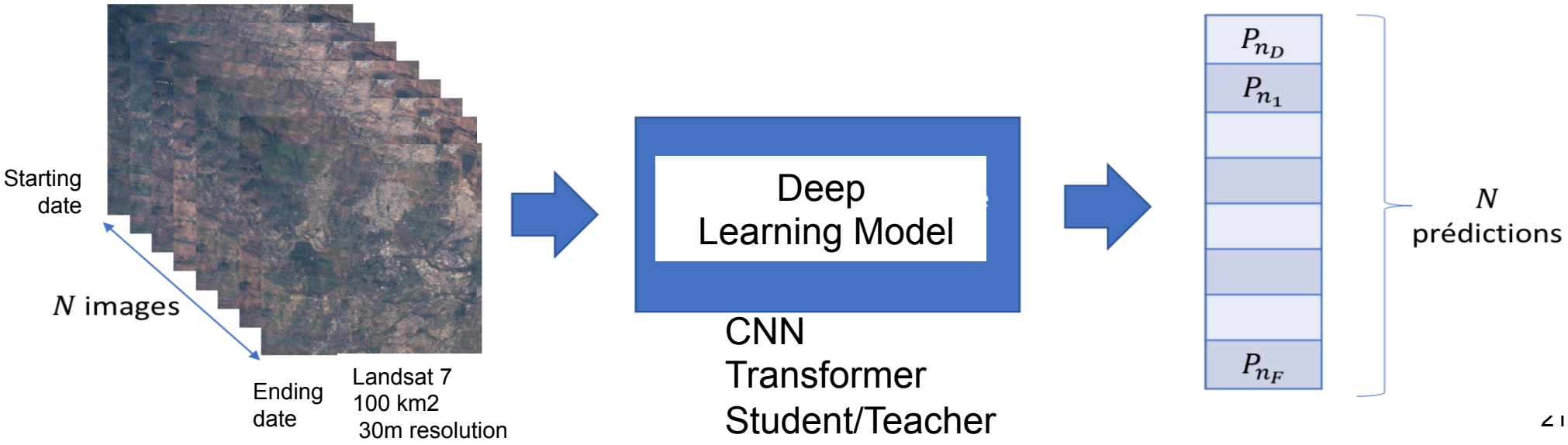
Economic wealth indicators



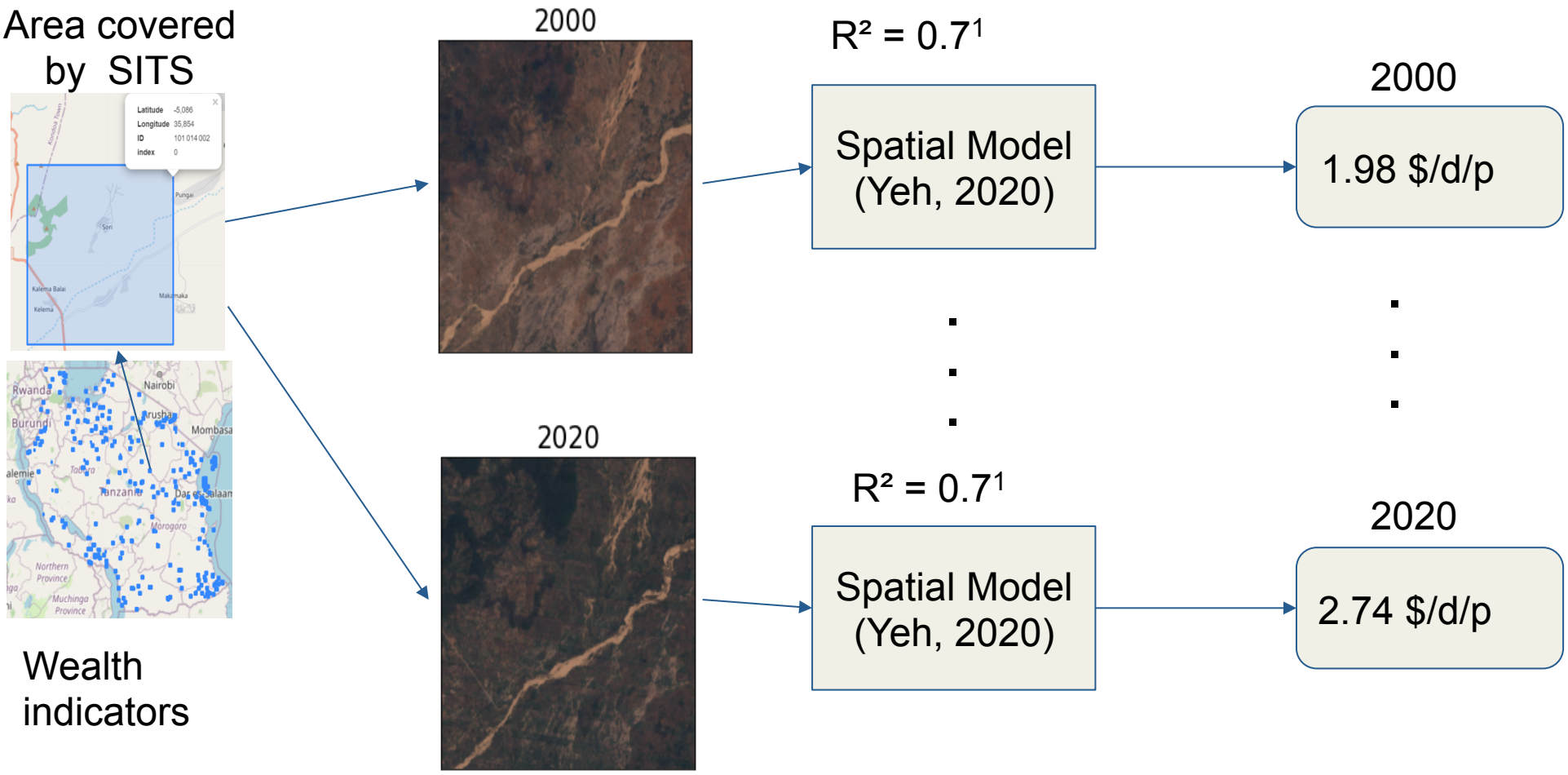
- Malawi : 3 surveys
- Ethiopia : 3 surveys
- Ouganda : 2 surveys
- Tanzania : 3 suveys

5800 ground truth data points as indicators

- **Sequences of images at different resolutions**



Poverty Evolution Prediction: A Naive Spatial Approach



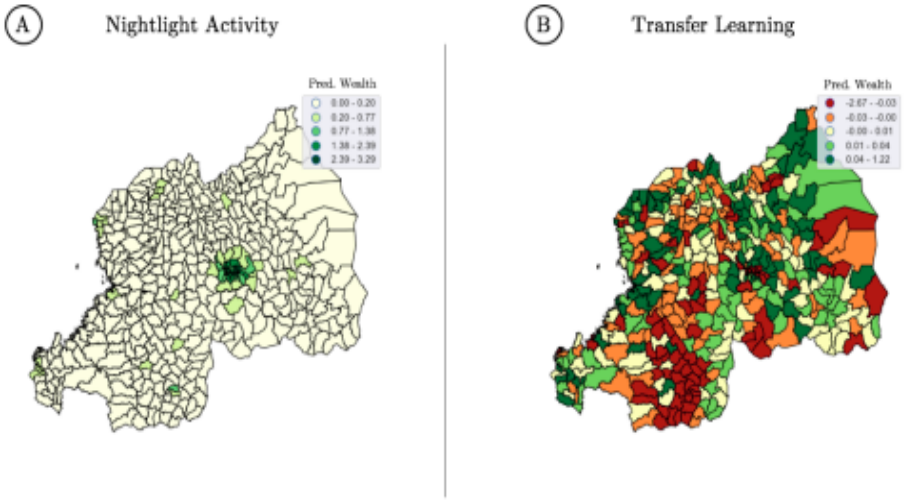
[1]C. Yeh, A. Perez, A. Driscoll, *et al.* "Using publicly available satellite imagery and deep learning to understand economic well-being in Africa." *Nat Commun* 11, 2583 (2020).

Poverty Evolution Prediction: Limited Results of the Naive Approach

[Kondmann et al., 2020][1]

	Mean Training R^2 2005	Test R^2 2010	Test R^2 2015
Transfer Learning & ResNet50	0.47	0.69	0.57
Mean Nightlights per cluster	0.52	0.74	0.61
ResNet50	0.55	0.43	0.36

Time	Change TL	Change NTL	Poverty Reduction	GDP Growth
2005 - 2010	0.00	0.03***	0.11	0.49
2010 - 2015	0.00	0.02***	0.07	0.44
2005 - 2015	0.00	0.05***	0.18	1.15

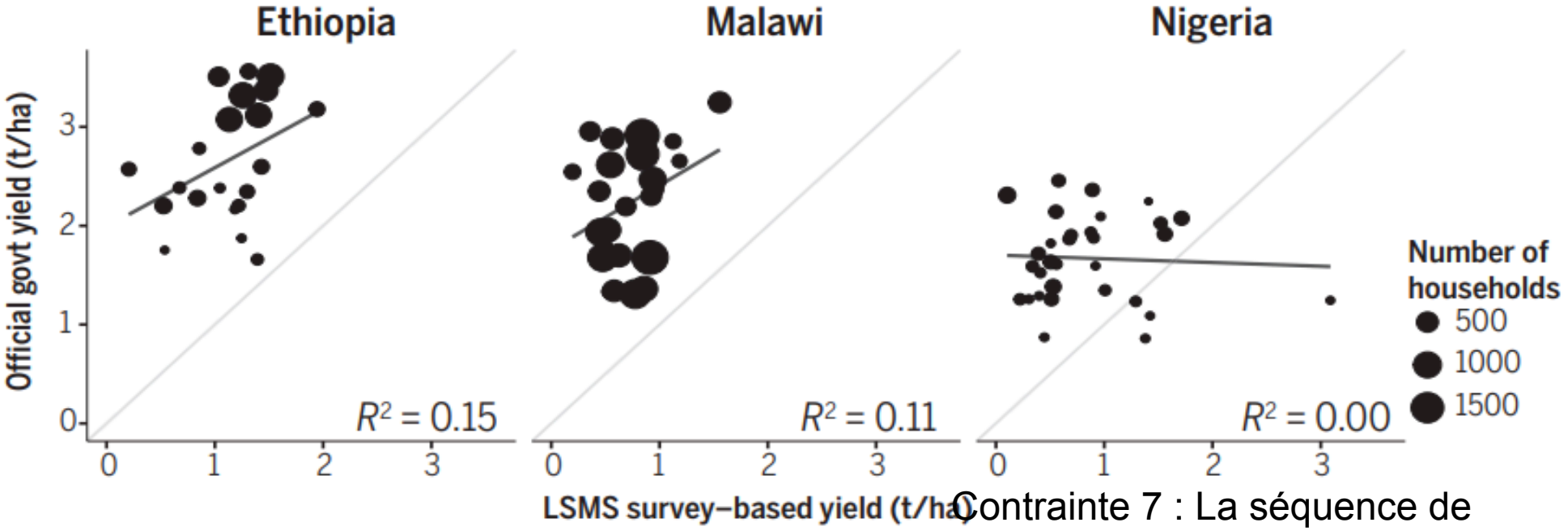


- Very heterogeneous evolution map
- **Consistency in time and space may not be preserved**

[1]L. Kondmann, et Zhu, X. X. "Measuring Changes in Poverty with Deep Learning and Satellite Images", *2020 ICLR Practical ML for Developing Countries Workshop*
[2]C. Yeh, A. Perez, A. Driscoll *et al.* "Using publicly available satellite imagery and deep learning to understand economic well-being in Africa." *2020 Nat Commun 11*.

Poverty Evolution Prediction: Naive Approach : Limited Results

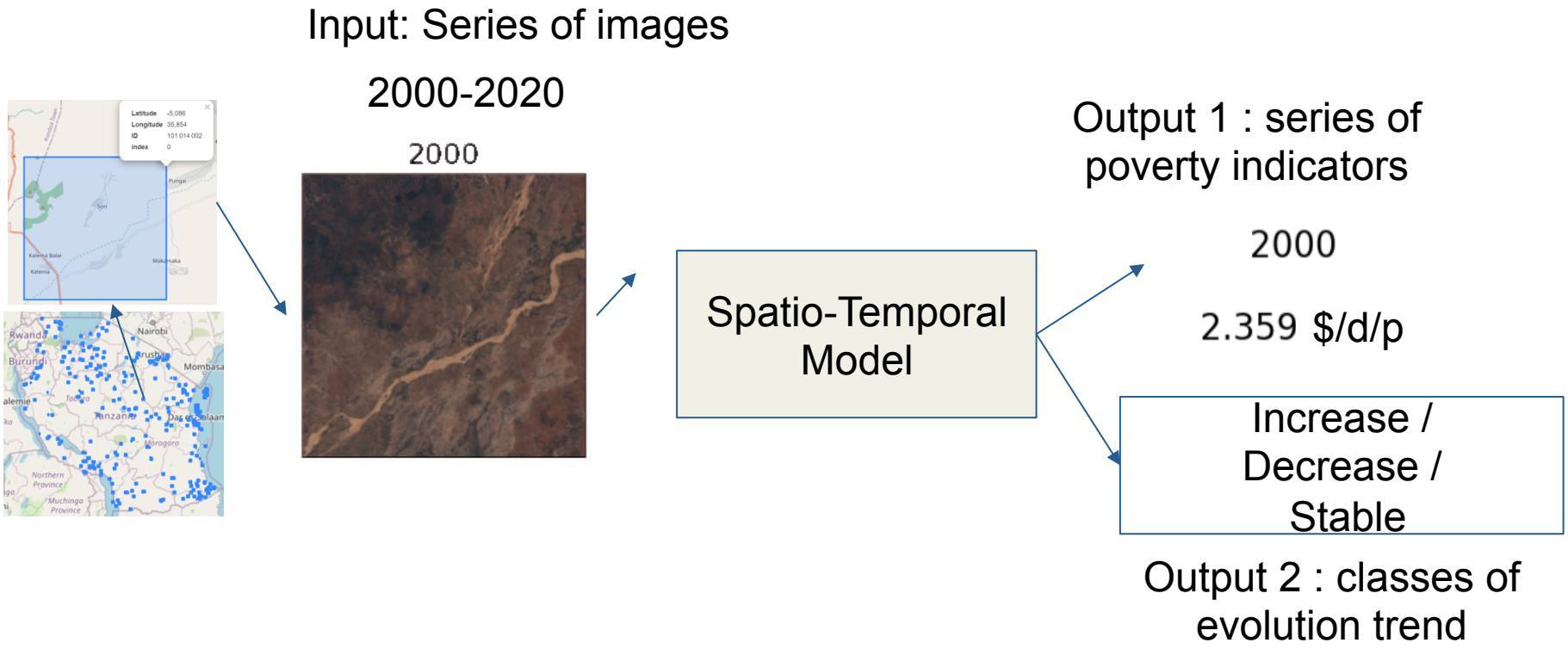
Burke et al, 2021[1]



Contrainte 7 : La séquence de pauvreté est altérée

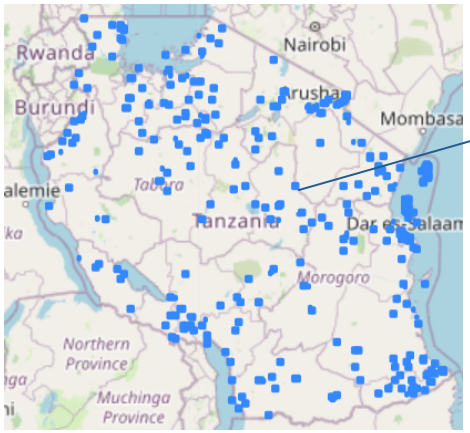
[1]Marshall Burke, Anne Driscoll, David B. Lobell, Stefano Ermon. “Using satellite imagery to understand and promote sustainable development”. *2021 Science*

Poverty Evolution Prediction using Transformers

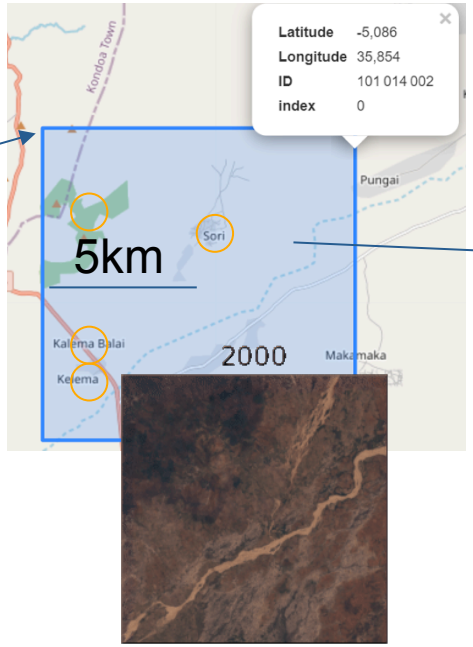


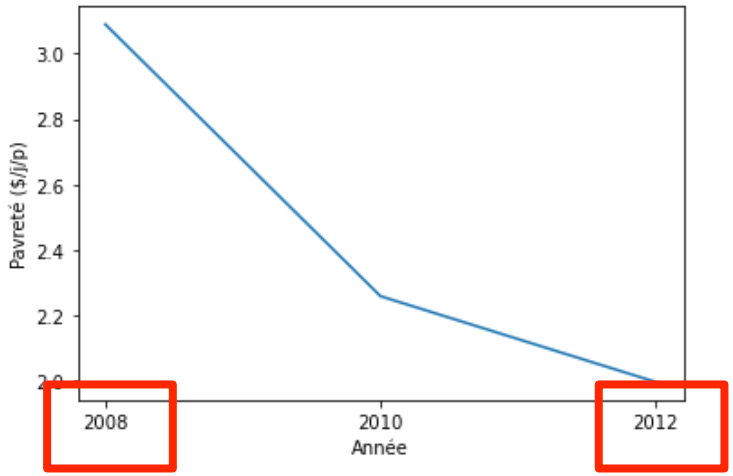
Constraint 1 : Avg Spatial Resolution (30m)
Constraint 2 : Avg Temporal Resolution (1 an)
Constraint 3 : Sensor Noise in the signal

Poverty Evolution Prediction: Constraints



409 points in Tanzania (LSMS ISA)



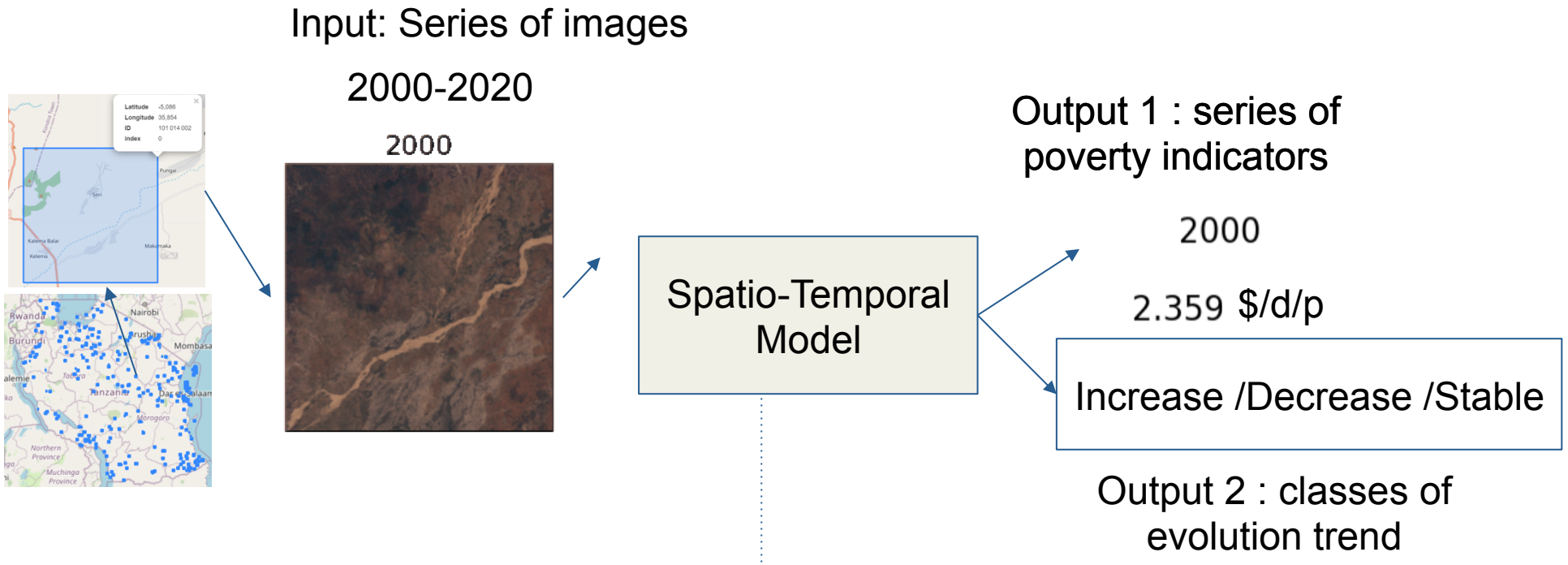


Constraint 4 : Data Sparsity

Constraint 5 : Randomization of the poverty indicator for anonymization

Constraint 6 : Relatively Small time window (few number of points)

Poverty Evolution Prediction: Need for intermediate training



Intermediate Training on NLI (proxy) Night Light Intensity

- Much more data available
- More accurate locations
- Longer time series
- *Presumably correlated with poverty evolution*

Challenges in Poverty Evolution Prediction from Satellite Images

1. **Neural architecture search and tuning is difficult, especially when training data is scarce in time/space**

2. **Finding the right proxy for pre-training is crucial**

3. **Finding the right resolution and aggregation level in time/space is challenging**

4. **Uncertainties are at every stage of the process**

Outline

I- Challenges in ML & data science pipelines

- Building the pipelines
- Preparing the data

II- Current Projects

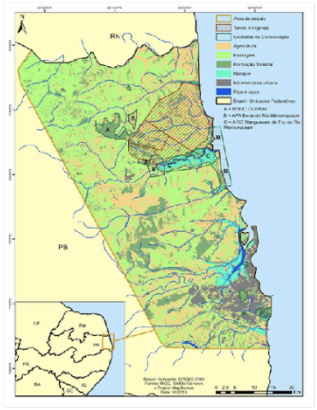
- Combining satellite imagery and socio-economic indicators to estimate poverty evolution in Africa
- Finding sustainable transition pathways in Nordeste

Finding Sustainable Pathways

https://ideal.ufpb.br/

IDEAL International Joint Lab

Area of study



Actors

- Administrators
- Site Managers
- Elected Representatives
- Exploitation owners
- Farmers
- Indigenous representatives

Finding Sustainable Pathways

https://ideal.ufpb.br

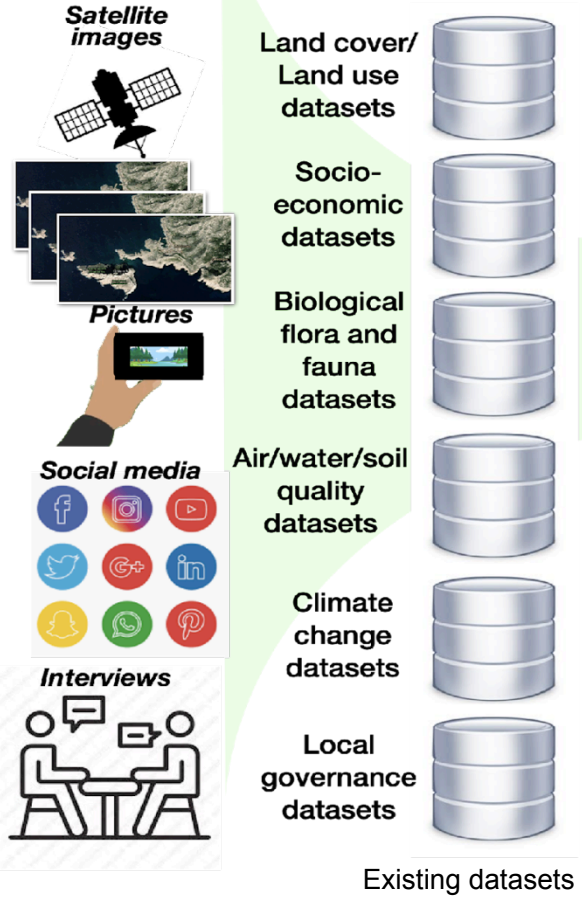
IDEAL International Joint Lab

Area of study

Actors

- Administrators
- Site Managers
- Elected Representatives
- Exploitation owners
- Farmers
- Indigenous representatives

Data Science & Engineering



Existing datasets

Finding Sustainable Pathways

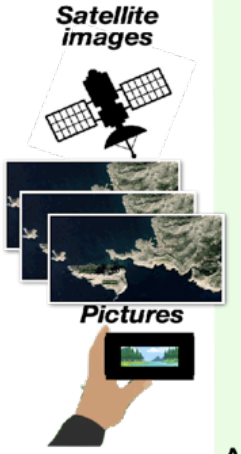
https://ideal.ufpb.br

Area of study

Actors

- Administrators
- Site Managers
- Elected Representatives
- Exploitation owners
- Farmers
- Indigenous representatives

Data Science & Engineering



- Land cover/ Land use datasets
- Socio-economic datasets
- Biological flora and fauna datasets
- Air/water/soil quality datasets
- Climate change datasets
- Local governance datasets

Existing datasets

Coviability Diagnosis

territorial scale

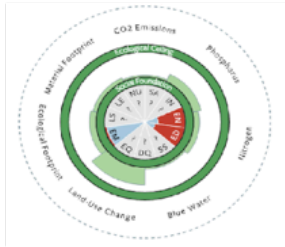
regional or departmental scale

local scale

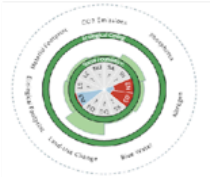
Multiple views as many as actors

Ideal Target

territorial scale



regional or departmental scale



local scale

Multiple co-specified targets

SHS Fieldwork with local populations

Finding Sustainable Pathways

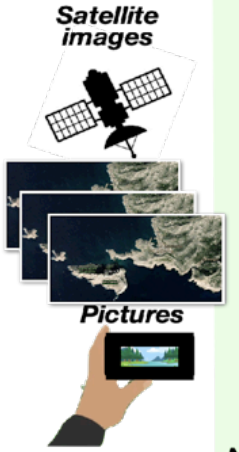
https://ideal.ufpb.br/

Area of study

Actors

- Administrators
- Site Managers
- Elected Representatives
- Exploitation owners
- Farmers
- Indigenous representatives

Data Science & Engineering



Land cover/ Land use datasets

Socio-economic datasets

Biological flora and fauna datasets

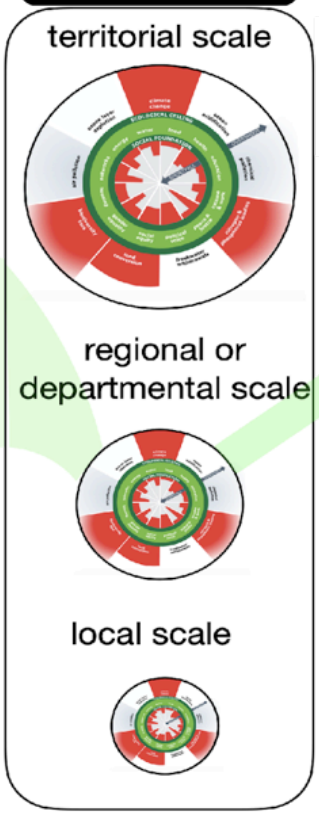
Air/water/soil quality datasets

Climate change datasets

Local governance datasets

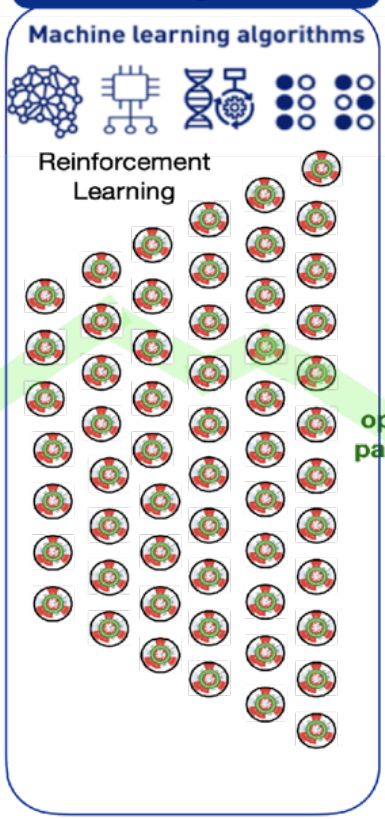
Existing datasets

Coviability Diagnosis

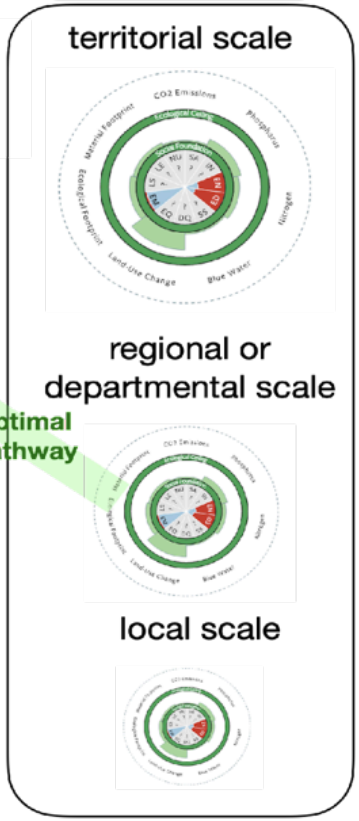


Multiple views as many as actors

Artificial Intelligence



Ideal Target

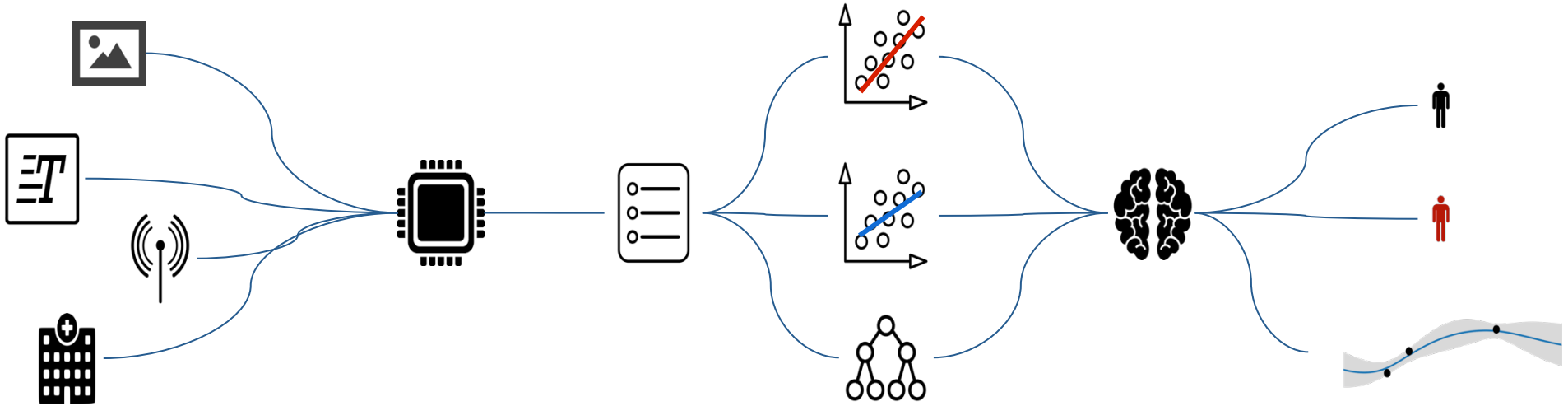


optimal pathway

Multiple co-specified targets

SHS Fieldwork with local populations

Challenges in Data and Decision Fusion



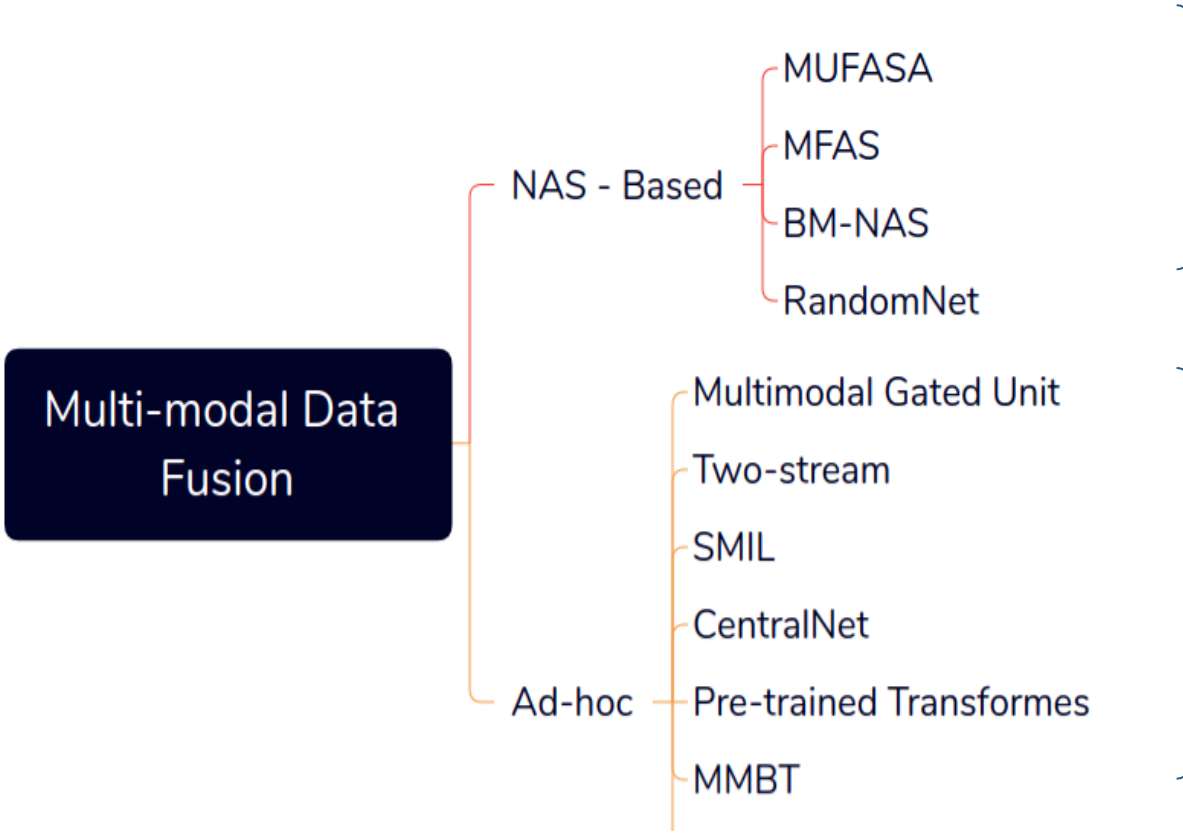
Multimodal data

Data Fusion

Decision Fusion

Prediction, Uncertainty quantification

Architectures for Multimodal Data Fusion



\+ Ideally will find the best architecture on each dataset and task
\- Very computationally expensive and slow
\- In practice not always manage to find the best architecture

\+ Computationally less expensive than NAS models
\+ Because of smaller time complexity allow to design and train more complex architectures
\- Need careful planning and manual adaptation for different tasks
\- Models designed for one task/ dataset may not transfer well to others

Concluding Remarks

- ML and data analytics provide efficient tools to help us answering many SD questions
- But crucially need principled data curation and prep
- ML for SD require **humans in the loop** and **UQ to provide actionable outputs**
- There are many opportunities for:
  - Managing and orchestrating human/machine resources
  - Proposing impactful ML applications for SDGs
  - Revisiting our methods & technologies **to serve SD**

Thanks!