Data-Driven Environmental System Analysis: Addressing Data Gap in Life Cycle Assessment by Using Artificial Intelligence

Bu Zhao, Ph.D.
Schmidt AI for Science Fellow

Process-Energy-Environmental Systems Engineering (PEESE)
Robert Frederick Smith School of Chemical and Biomolecular Engineering
Cornell University
Ithaca, NY

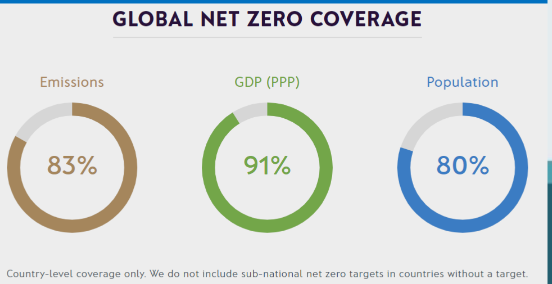
www.peese.org

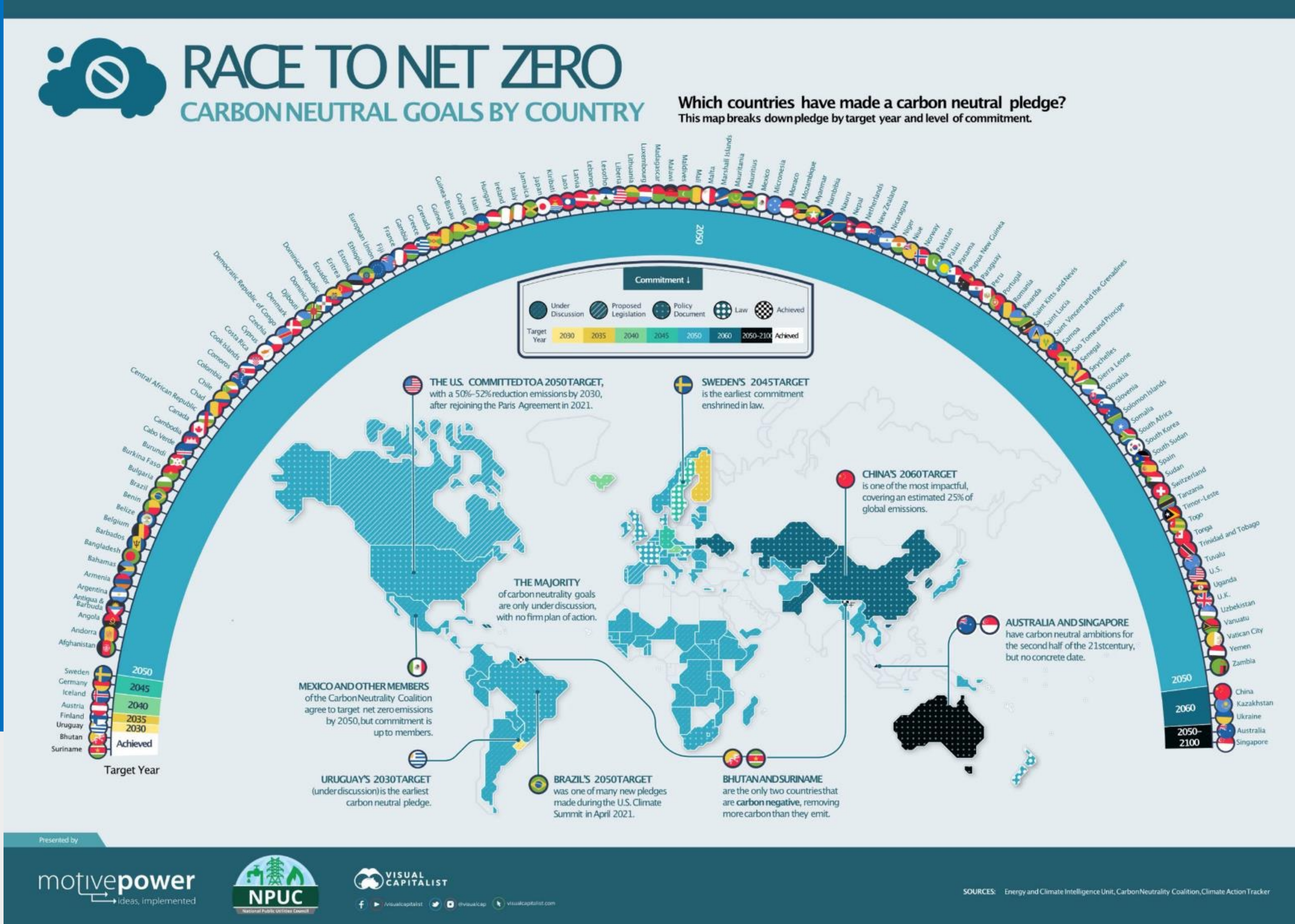
BEST 402 INDUSTRIAL ECOLOGY **11/09/2023**

TABLE OF CONTENT

139 countries announced carbon neutrality goals

by Nov. 24, 2022
https://zerotracker.net/





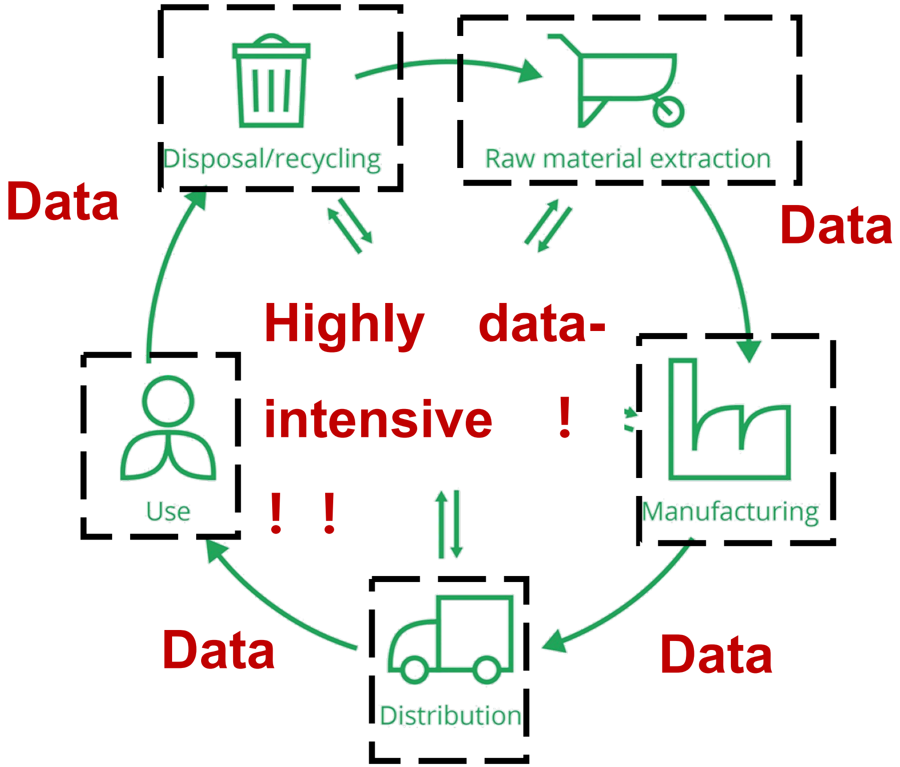
Life cycle thinking and carbon footprint

- **Carbon footprint** labels become "standard" for many products

UK USA South Korea Canada

Japan France North America

- Carbon footprint considers both ***direct and indirect*** GHG emissions in the product's life cycle



- ***Life Cycle Assessment*** is a standard method to assess environmental impacts associated with a product during its ***entire life cycle***.

Two types of data in LCA

LIFE CYCLE INVENTORY，LCI

- **Life cycle inventory (LCI)** is the methodology step that involves creating an inventory of input and output flows for a product system.
  - **Unit Process Data**

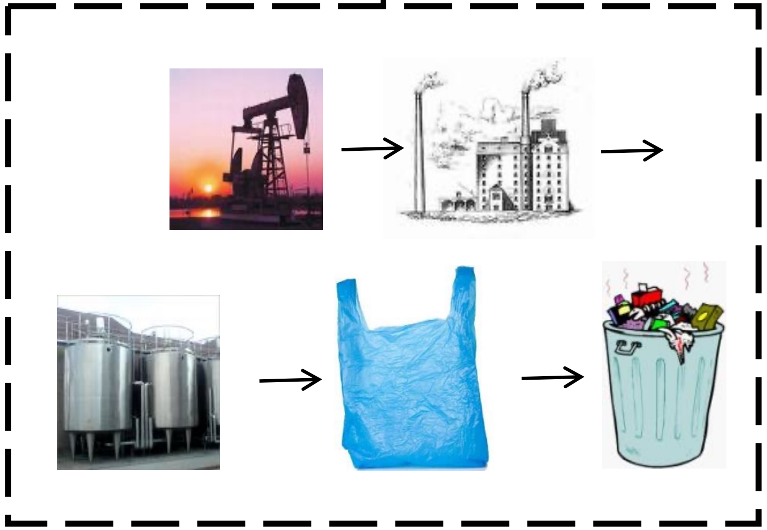
LIFE CYCLE IMPACT ASSESSMENT，LCIA

- **Life cycle impact assessment (LCIA)** is a step for evaluating the potential environmental impacts by converting the LCI results into specific impact indicators.
  - **Characterization factor**

AND

	Plastic film (kg)
Polyethylene	1.02 kg
Electricity	0.66 kWh
Wastewater	27 L
CO_2	1.42 kg
Methane	0.08 kg
…	…

…



CO_2

CH_4

…

$$EI = \sum LCI \times CF$$

Global warming potential (GWP)

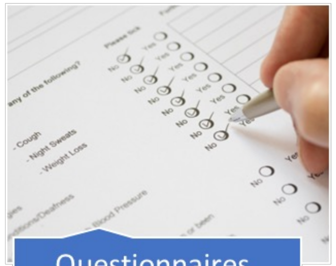
GWP $= CO_2 \times 1 + CH_4 \times 28 + \cdots$

Data gap in LCA studies

Current data collection methods:

Onsite investigation

Laboratory test



Questionnaires, surveys

Literature review

Expensive &Time consuming

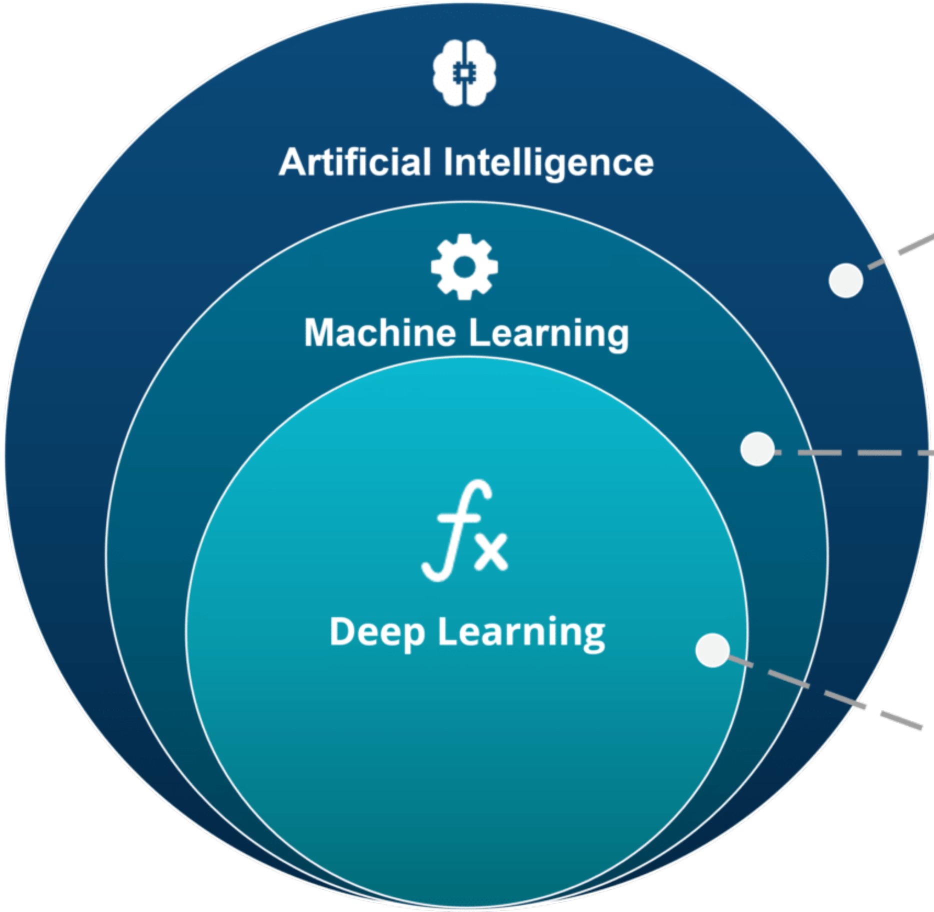
Missing data in LCI & LCIA

More Efficient Method

- **Data Availability**
- **Data Quality**
- Can we build a ***data-driven computational framework*** for estimating missing LCA data based on the existing data?

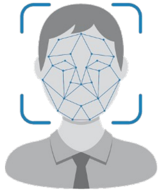
https://www.researchgate.net/publication/256378751_JRC_Reference_Reports_The_International_Reference_Life_Cycle_Data_System_ILCD_Handbook/figures?lo=1

Artificial intelligence provides additional insights



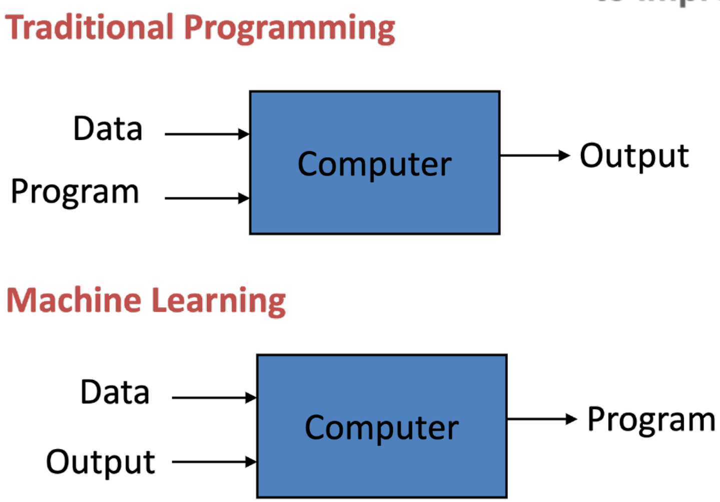
ARTIFICIAL INTELLIGENCE

A technique which enables machines to mimic human behaviour

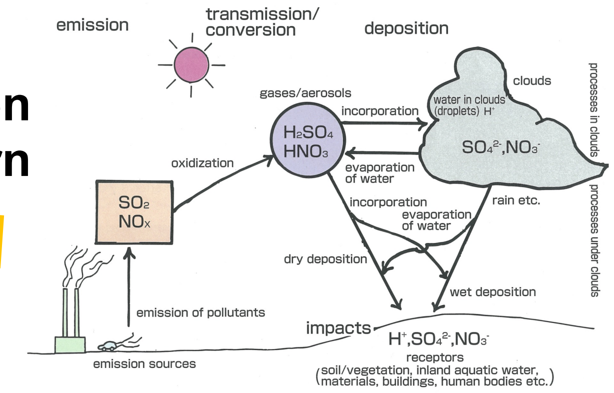


MACHINE LEARNING

Subset of AI technique which use statistical methods to enable machines to improve with experience



Hidden pattern



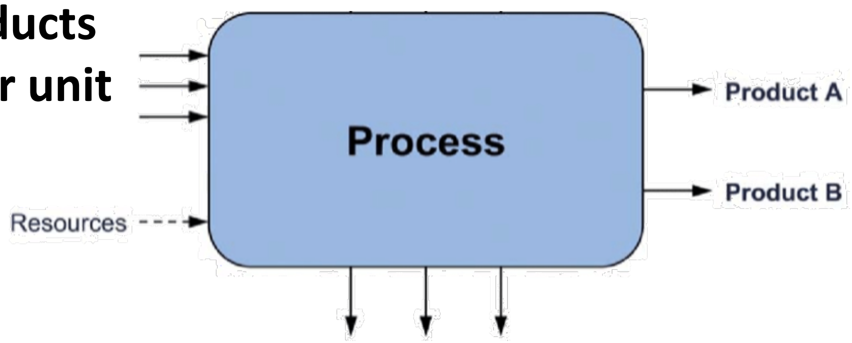
https://www.edureka.co/blog/ai-vs-machine-learning-vs-deep-learning/

TABLE OF CONTENT

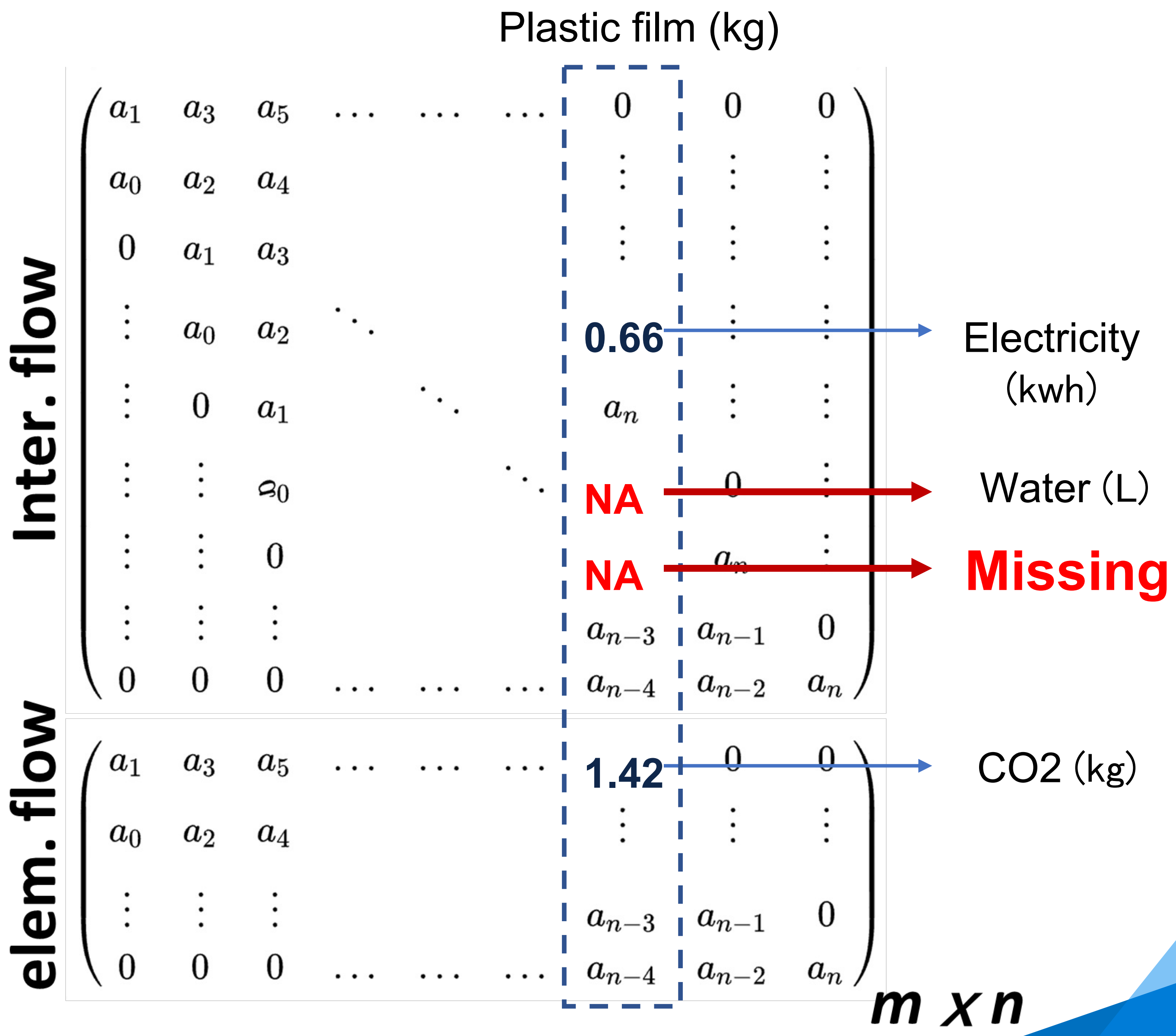
Life Cycle Inventory (LCI)

- Unit process dataset (ecoinvent database)

Input products from other unit processes

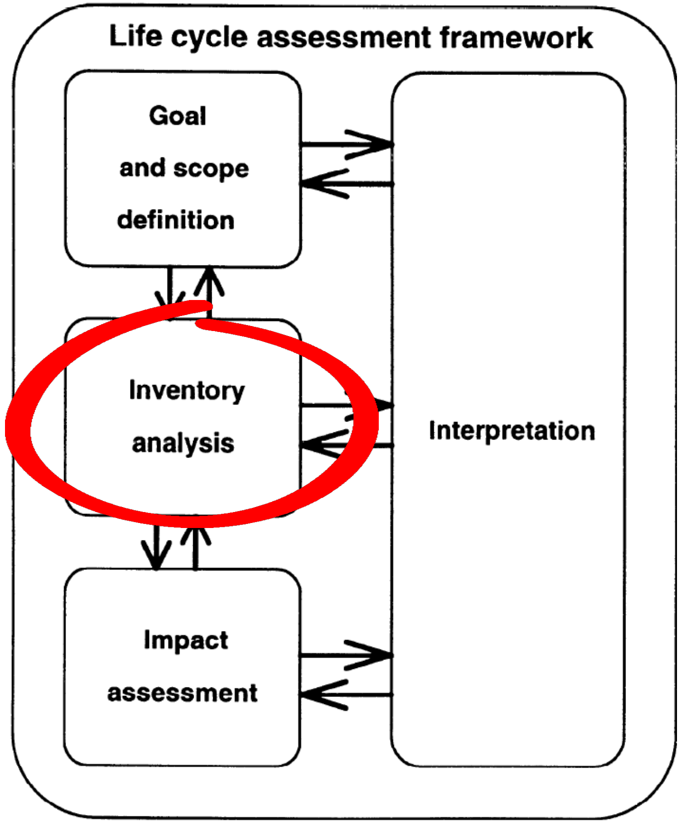


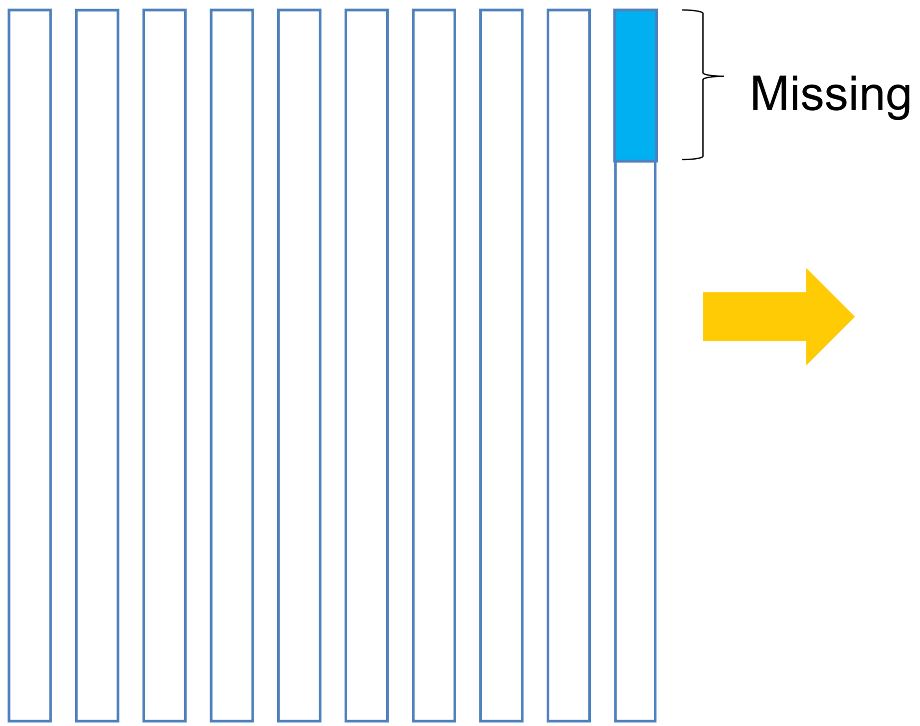
Wastes and emissions



Research question and goal

Research goal: advance LCA data compilation by developing a computational framework for estimating missing LCA data based on the existing data.



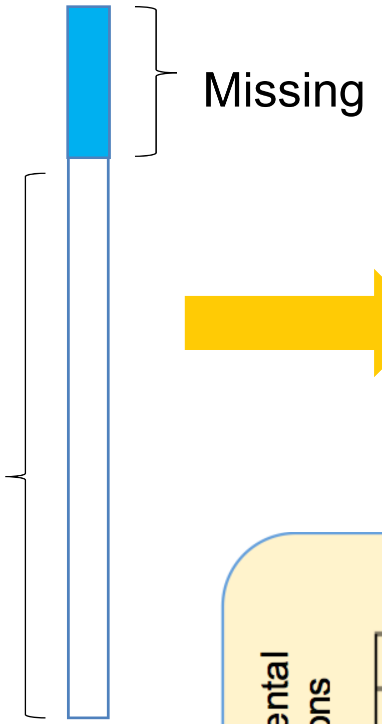


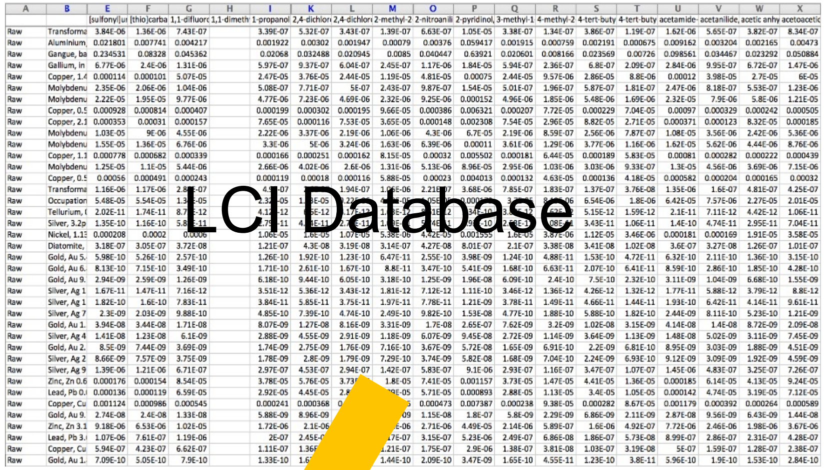
Partially complete unit process database

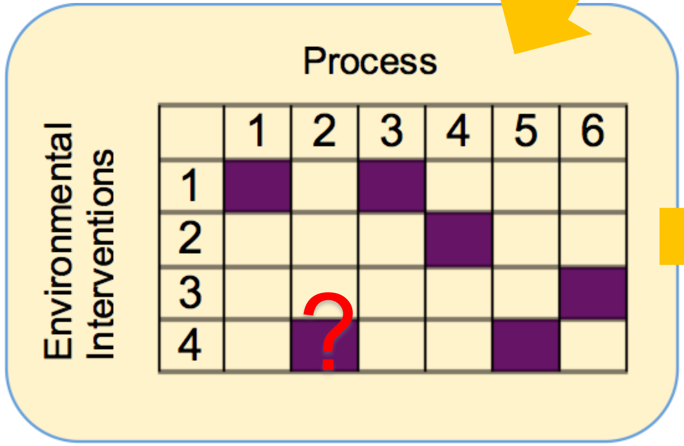
Two approaches:

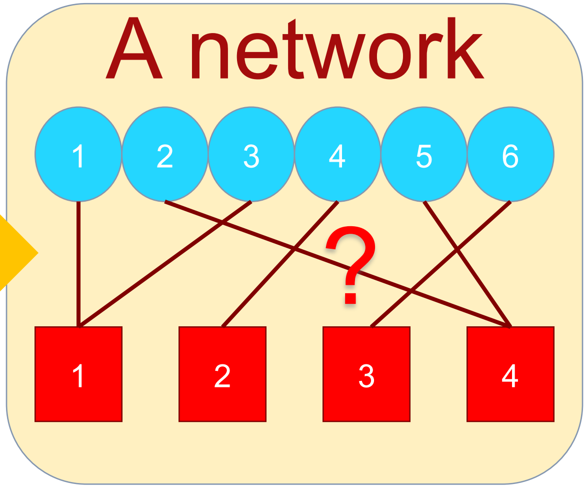
1. **Similarity-based link prediction**
2. **Decision-tree based supervised learning**

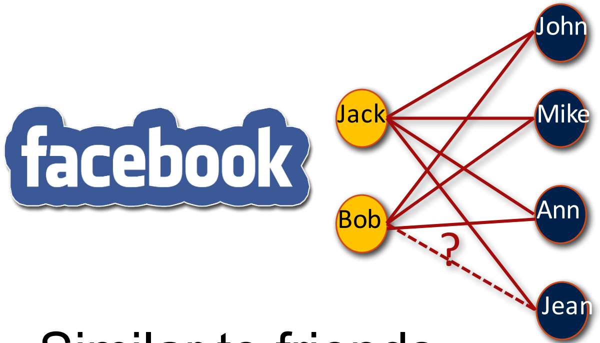
Similarity-based link prediction method











Similar to friends recommendation in social media

Estimating missing data in LCI database = predict missing links in a network

- Hou, P., Cai, J., Qu, S., & Xu, M. (2018). Estimating missing unit process data in life cycle assessment using a similarity-based approach. *Environmental science & technology*, *52*(9), 5259-5267.

Algorithm and procedure

1. Calculate similarity:

Minkowski distance：

$$d_{ij} = \left(\sum_{t=1}^{m-p} |a_{ti} - a_{tj}|^q \right)^{1/q}$$

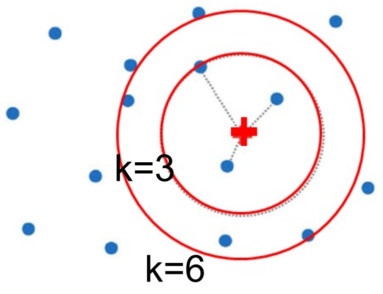
Similarity：

$$s_{ij} = \frac{1}{d_{ij} + 1}$$

2. Estimate missing data:

The weighted mean of k most similar processes.

$$e_{tj} = \frac{\sum_{i=1}^{k} a_{ti} s_{ij}}{\sum_{i=1}^{k} s_{ij}}$$



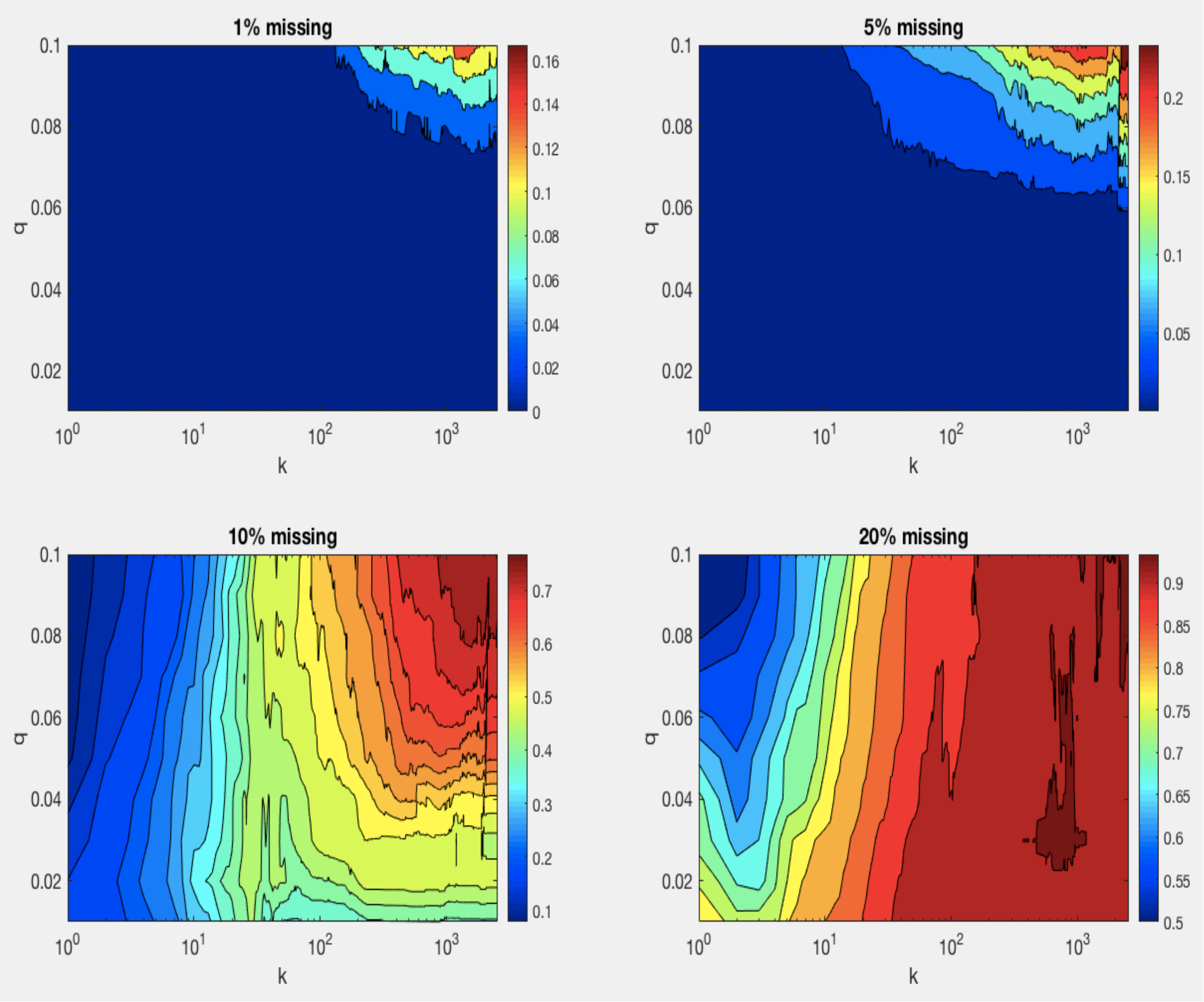
E1 = (2.0*s1+4.1*s2+9.8*s3)/(s1+s2+s3)

	Target Process	Process 1	Process 2	Process 3	…
Input 1	**E1**	2.0	4.1	9.8	…
Input 2	2.0	3.0	7.1	7.4	…
Input 3		1.0	5.9	9.1	…
Input 4	1.0	0.2	5.4	4.9	…
Output 1	0.5	0.4	1.8	6.1	…
Output 2		0.6	6.6	3.7	…
Output 3	2.0	2.0	7.4	0.2	…
Output 4	3.0	1.0	1.4	0.2	…
…	…	…	…	…	…

Note: Ranked in descending order of similarity

The best parameters are selected using the leave-one-out-cross-validation (LOOCV) on training set

Results - MPEs with different data missing



- When fewer data are missing (**1% and 5%**), the estimation MPEs are distributed in relatively narrow ranges with **very high accuracy**;
- When more data are missing (**10%**), the distribution of MPEs becomes much broader and model performance can **vary greatly between different unit processes**;
- When missing data exceeds a certain level (**20%**), the known information is not enough to find the true similar processes and thus the method is **hard to estimate those missing data**.
- Need to find another flexible method to **solve the situation when more data are missing**

Hypothesis for the second model

11,332 columns (unit processes)

13,201 rows (inter. and elem. flows)

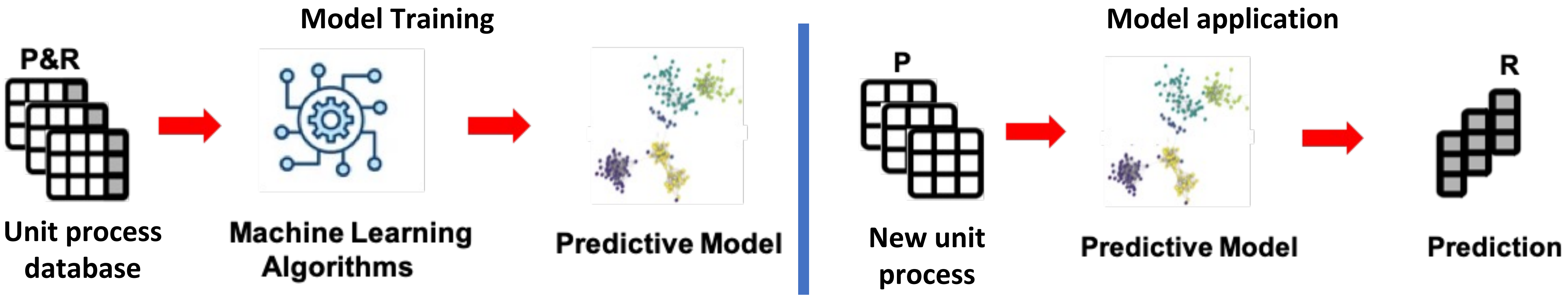
- Flows and are somewhat correlated with each other

$$
\overbrace{\hphantom{\begin{pmatrix} a_1 & a_3 & a_5 & \cdots & \cdots & \cdots & 0 & 0 & 0 \end{pmatrix}}}^{\text{Unit Process}}
$$

$$
\begin{array}{cc}
\text{Inter. flow} &
\begin{pmatrix}
a_1 & a_3 & a_5 & \cdots & \cdots & \cdots & 0 & 0 & 0 \\
a_0 & a_2 & a_4 & & & & \vdots & \vdots & \vdots \\
0 & a_1 & a_3 & & & & \vdots & \vdots & \vdots \\
\vdots & a_0 & a_2 & \ddots & & & 0 & \vdots & \vdots \\
\vdots & 0 & a_1 & & \ddots & & a_n & \vdots & \vdots \\
\vdots & \vdots & a_0 & & & \ddots & a_{n-1} & 0 & \vdots \\
\vdots & \vdots & 0 & & & & a_{n-2} & a_n & \vdots \\
\vdots & \vdots & \vdots & & & & a_{n-3} & a_{n-1} & 0 \\
0 & 0 & 0 & \cdots & \cdots & \cdots & a_{n-4} & a_{n-2} & a_n
\end{pmatrix} \\
\text{elem. flow} &
\begin{pmatrix}
a_1 & a_3 & a_5 & \cdots & \cdots & \cdots & 0 & 0 & 0 \\
a_0 & a_2 & a_4 & & & & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & & & & a_{n-3} & a_{n-1} & 0 \\
0 & 0 & 0 & \cdots & \cdots & \cdots & a_{n-4} & a_{n-2} & a_n
\end{pmatrix}
\end{array}
$$

m x n

Supervised learning approach



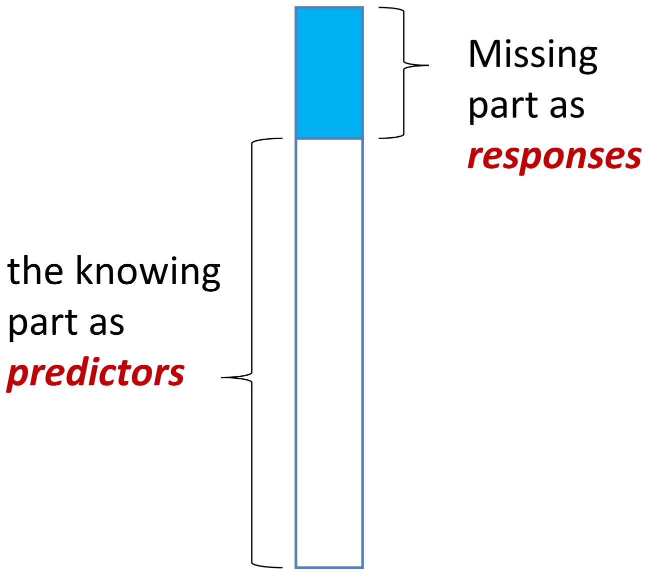


Table. Model performance on 10 test processes with 10% of data missing

	Classification Task		*Regression Task*		
Algorithms	*Misclassification rate*	*Computational Time*	R^2	*MPE*	*Computational Time*
XGBoost	0.72%	48 minutes	0.771	13.66%	6 minutes

Results - MPEs with different data missing

Table. R^2 and MPEs with different percentages of data missing

Percentage of missing data	*1%*	*5%*	*10%*	*20%*	*30%*	*50%*	*70%*
XGBoost, R^2	**0.751**	**0.734**	**0.730**	**0.713**	**0.623**	**0.487**	**0.272**
XGBoost, MPE	**15.24%**	**17.01%**	**17.74%**	**21.36%**	**38.47%**	**54.24%**	**74.79%**
RF, R^2	**0.541**	**0.517**	**0.511**	**0.495**	**0.432**	**0.362**	**0.215**
RF, MPE	**46.91%**	**50.96%**	**51.41%**	**55.43%**	**62.68%**	**73.25%**	**78.91%**

Summary

PEESE

This study demonstrates the promising potential of using computational approaches for LCI data compilation.

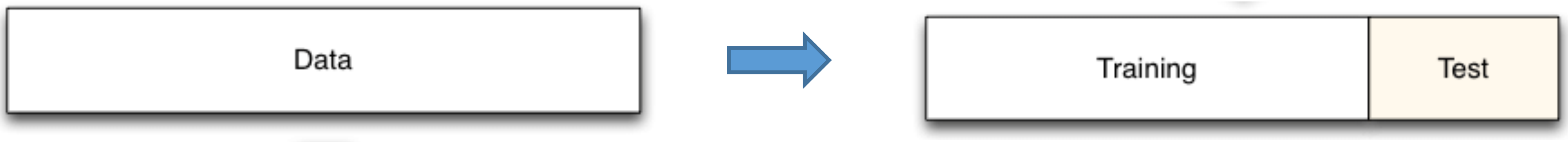
1
- This method does not intend to replace primary data collection, but is ***a complementary approach*** when primary data are not available.

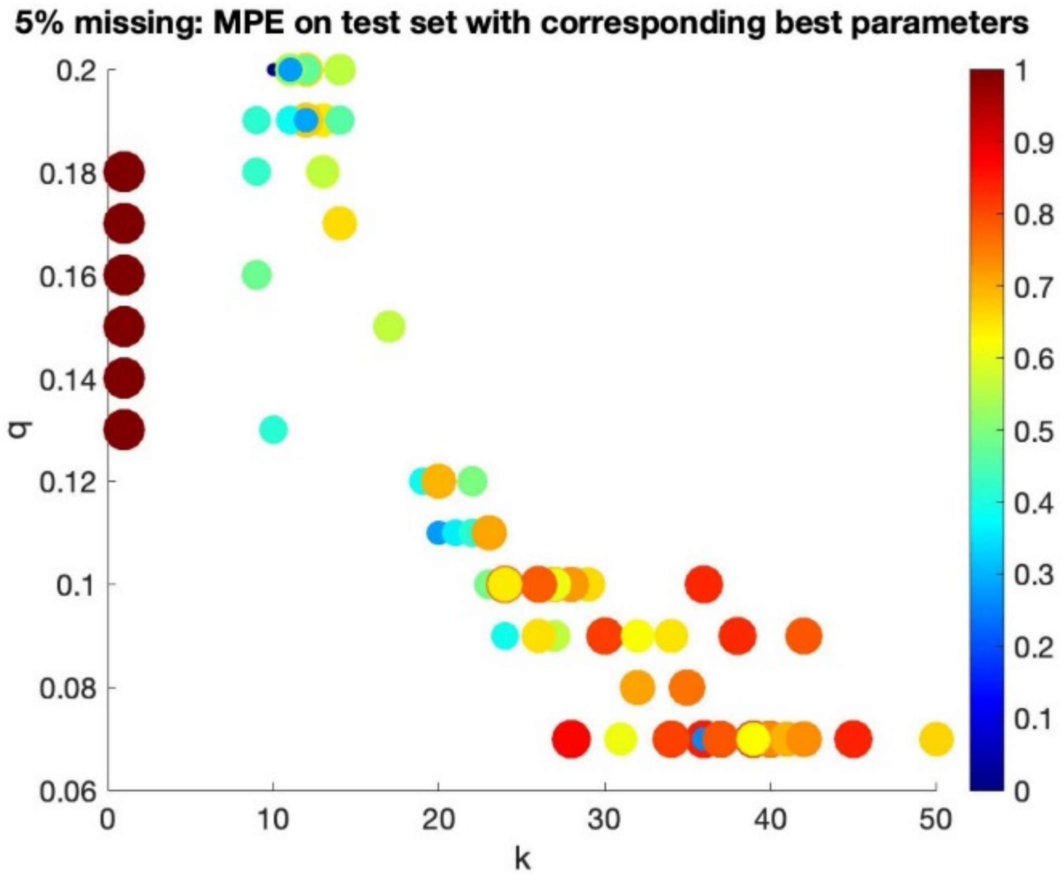
2
- This method can be used to ***upgrade the data quality*** for existing database.

3
- This method can be used to ***estimate the incomplete data for a new database*** based on part of its known data.

TABLE OF CONTENT

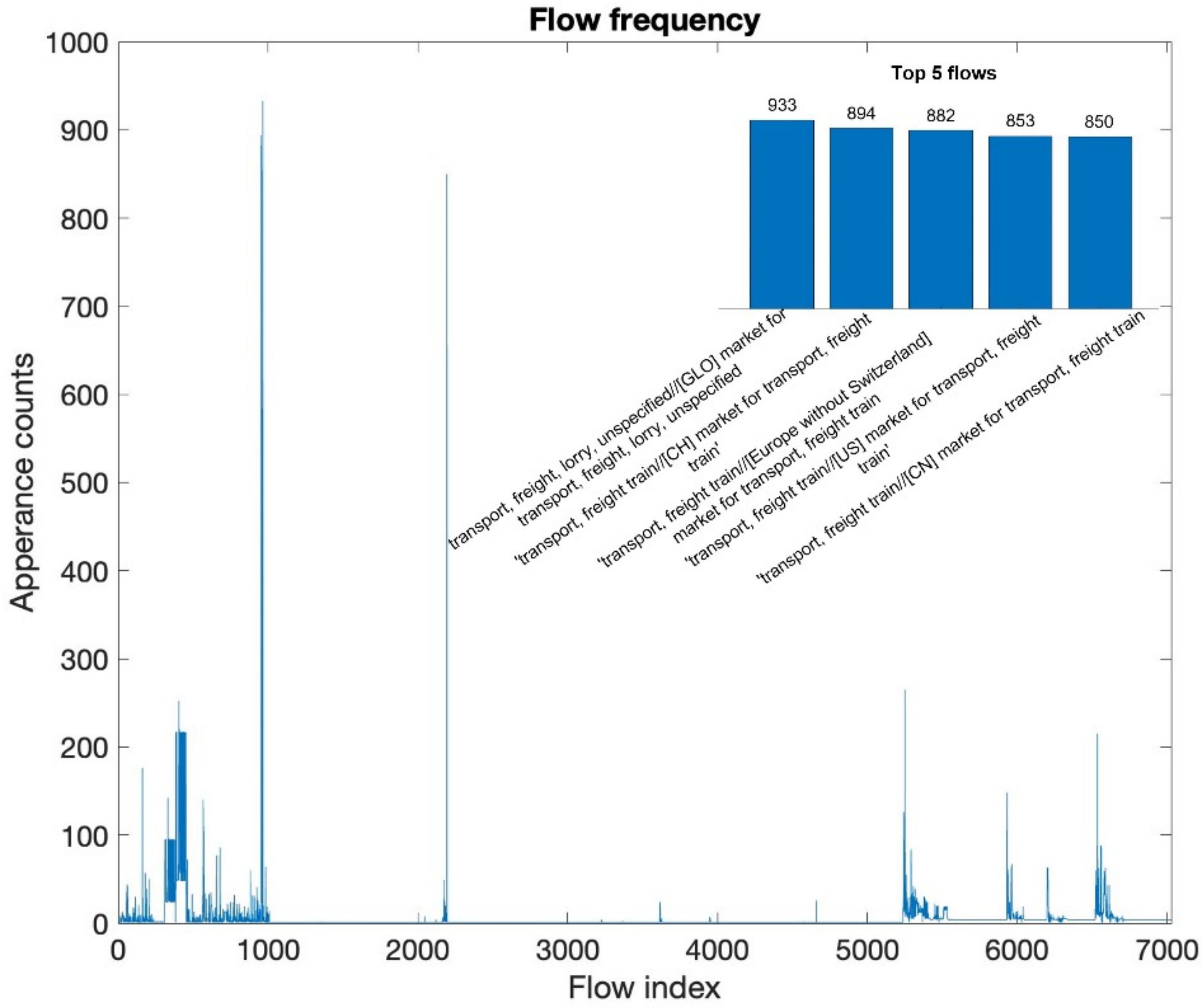
Variation due to random train-test data splits





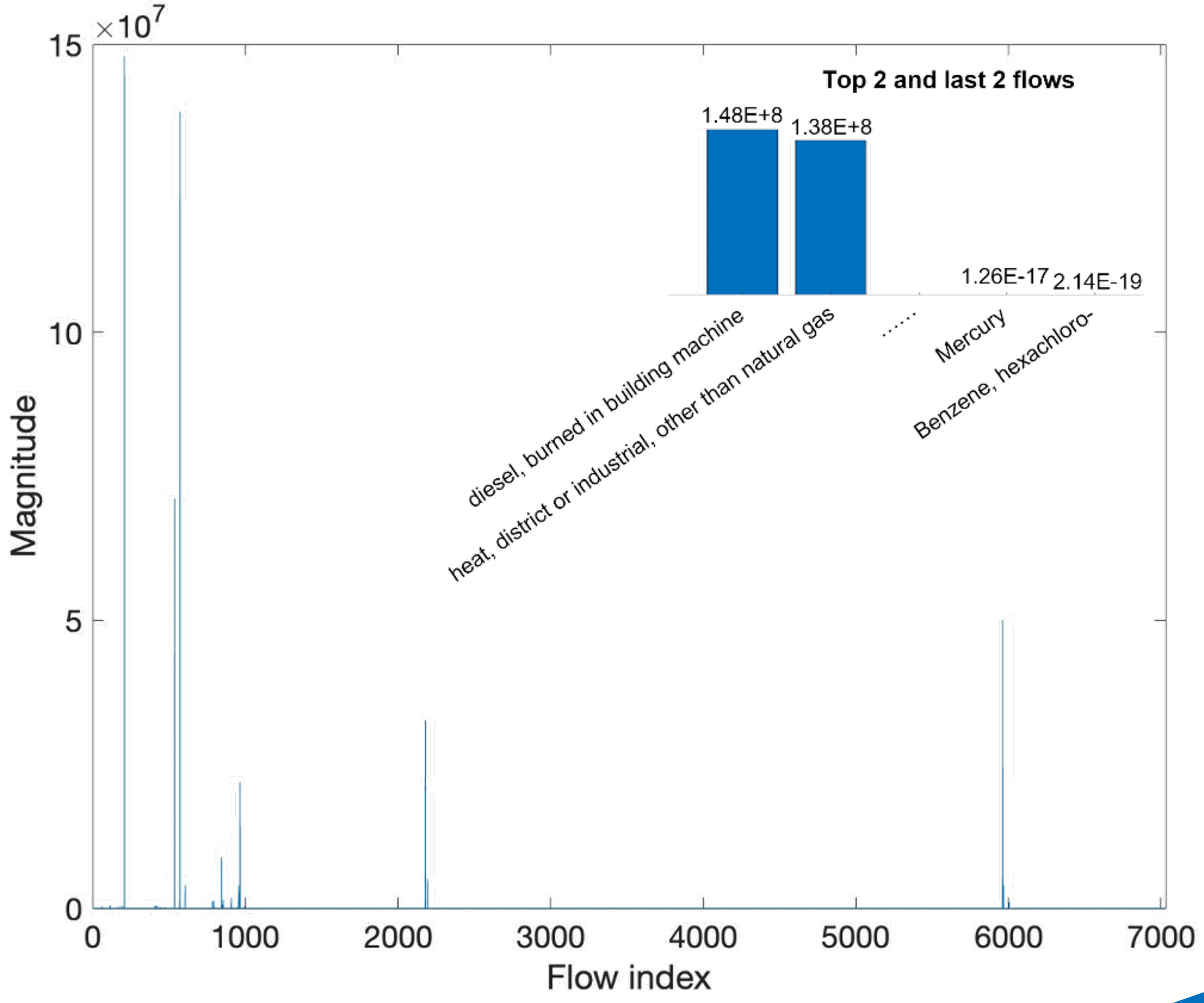
Data sparsity and imbalance

- The Ecoinvent UPR database is a **sparse matrix**, in which only **0.24%** of entries are nonzero.
- The **top 20%** of flows with the highest appearance accounted for **80% of the total non-zero flows**.



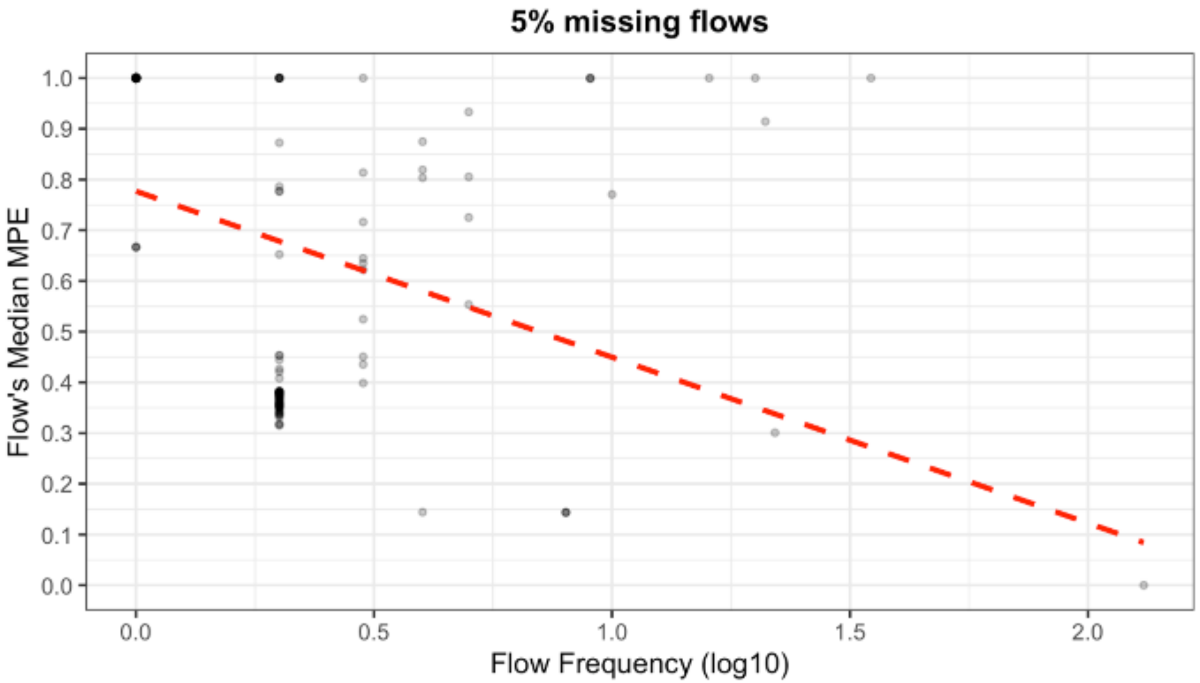
Data magnitudes

- The UPR database represents the underlying technology network which has clear physical and chemical meanings.
- Process: Transport freight train （**ton. km**)
- Flow: CO_2 (**kg**)

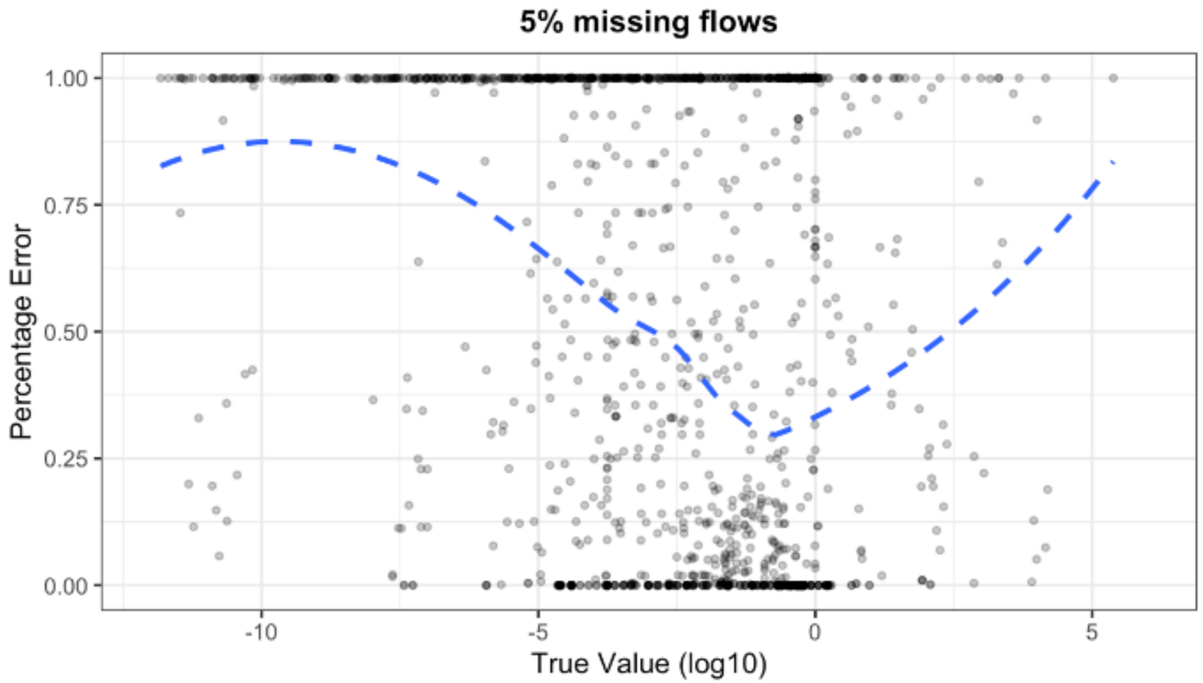


Variation due to imbalance of data

- The number of appearances had a **positive** impact on the model's performance
- **"U" shape** in the scatter plot between the magnitude of this flow and its MPE



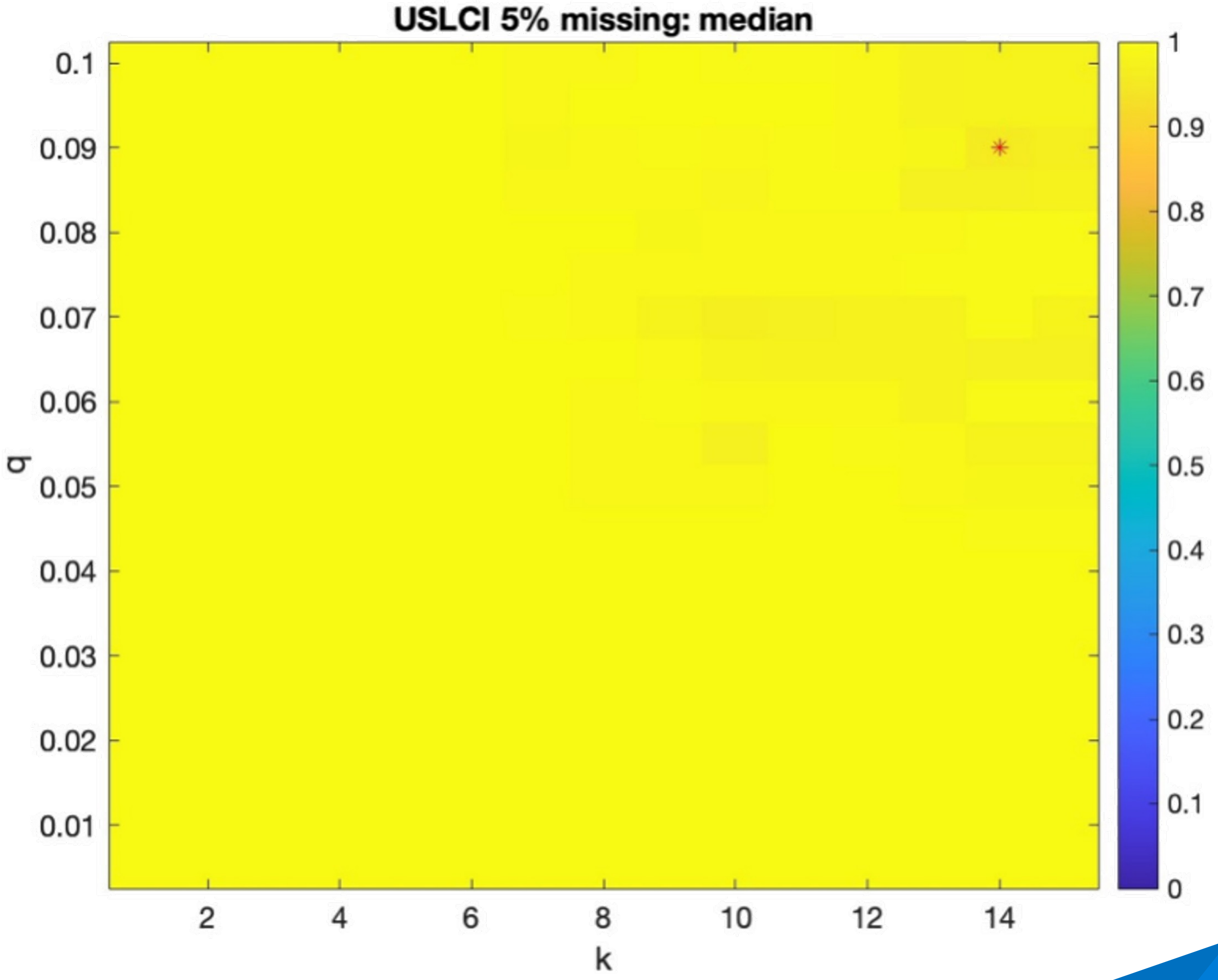
(a)



(b)

Model generalizability on different databases

- **U.S. Life Cycle Inventory (USLCI) Database**
- 4074 flows and 638 processes
- 7 times smaller than Ecoinvent
- Without information for other countries and regions



Summary

High variabilities in the existing machine learning methods for LCA studies due to the data and model selections.

1

- Data integration and data fusion from multiple sources are important for more accurate and less biased estimations.

2

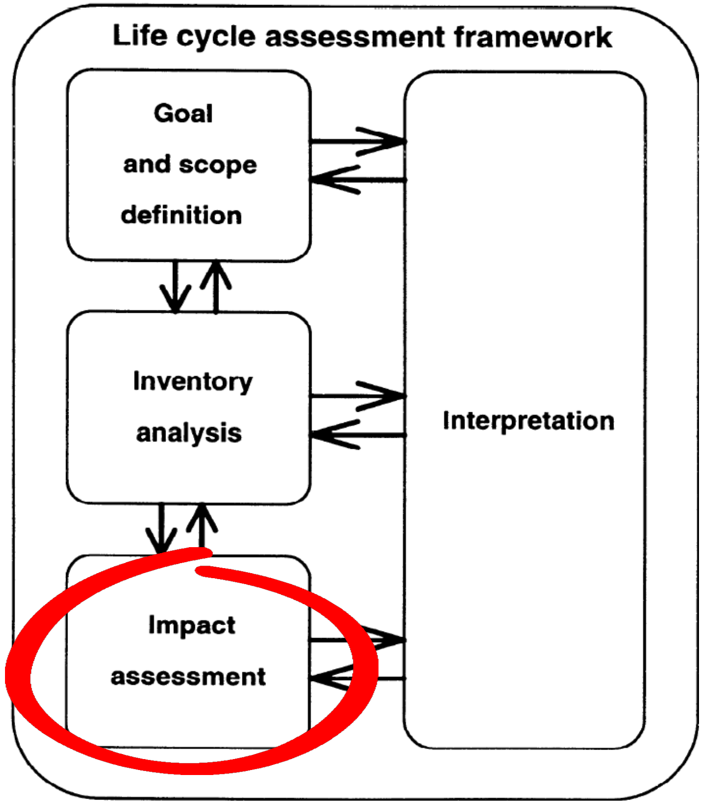
- Acknowledge the potential variation due to the randomness of the data.

3

- The trade-off between the “physical meaning” of the data and applying needed mathematical operations.

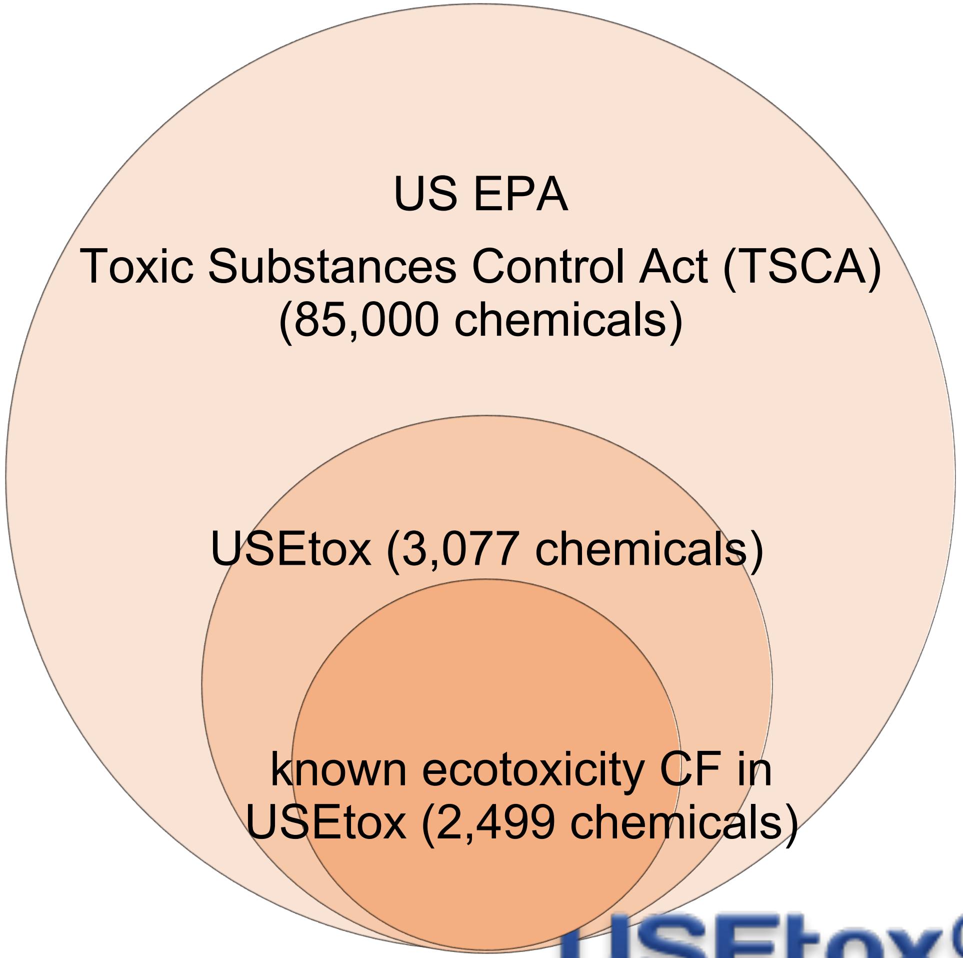
TABLE OF CONTENT

Characterization Factor in LCIA

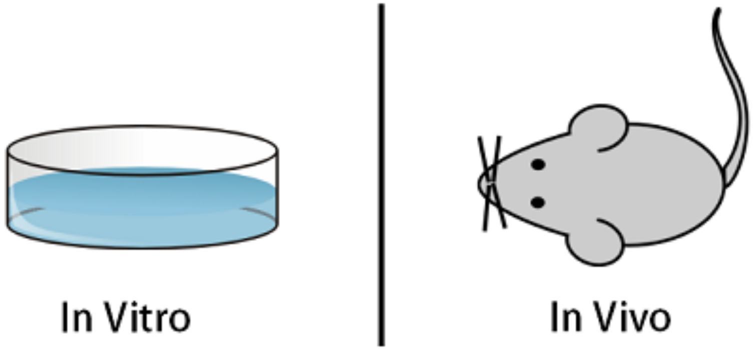


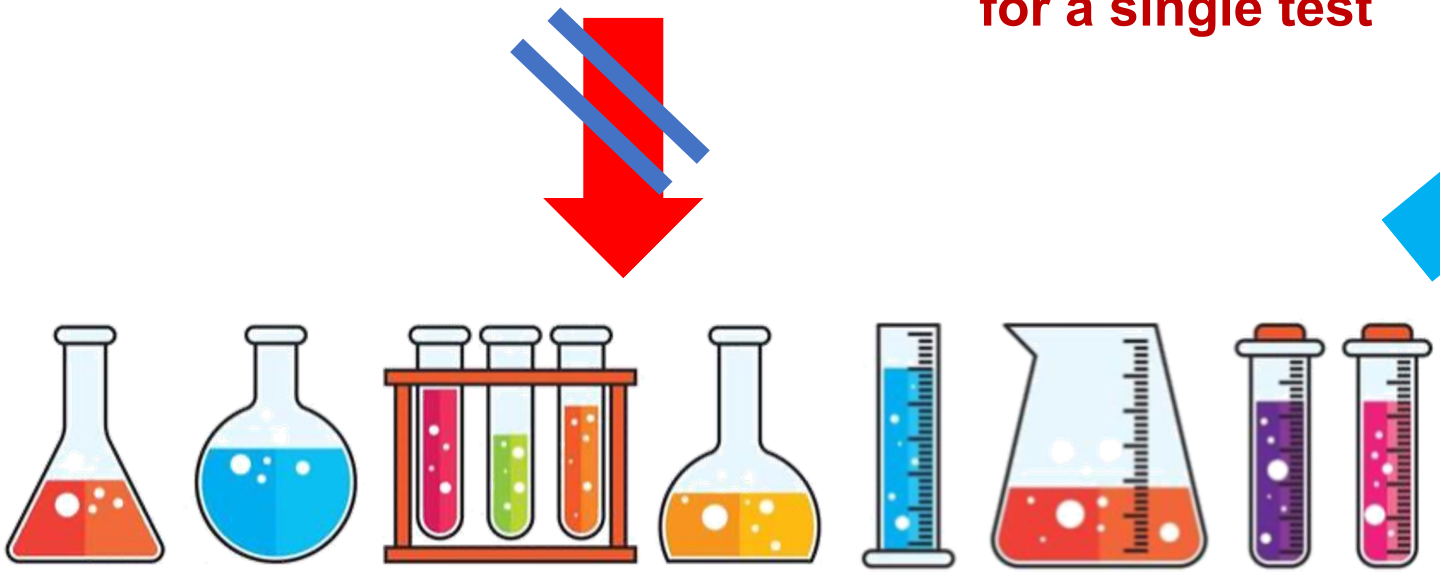
Life cycle impact assessment (LCIA) – quantifying the impacts of chemicals and other contaminants.

USEtox v2.0 provides **ecotoxicity** and human toxicity characterization factors (CF)

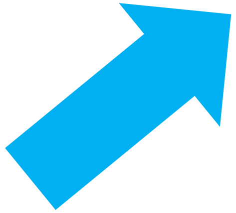


Why missing characterization factors





- **Could last for several months**
- **Could cost for \$8,000 to \$20,000 for a single test**



ECOSAR (Ecological Structure-Activity Relationship)

Organism	End Pt	mg/L (ppm)
Fish	LC50	0.835
Fish	LC50	1.773
Daphnid	LC50	1.275
Green Algae	EC50	0.344
Fish	ChV	0.034
Daphnid	ChV	0.381
Green Algae	ChV	0.225
Fish (SW)	LC50	1.065
Mysid	LC50	0.253
Fish (SW)	ChV	0.256
Mysid (SW)	ChV	0.053
Earthworm	LC50	461.663 *

HC50

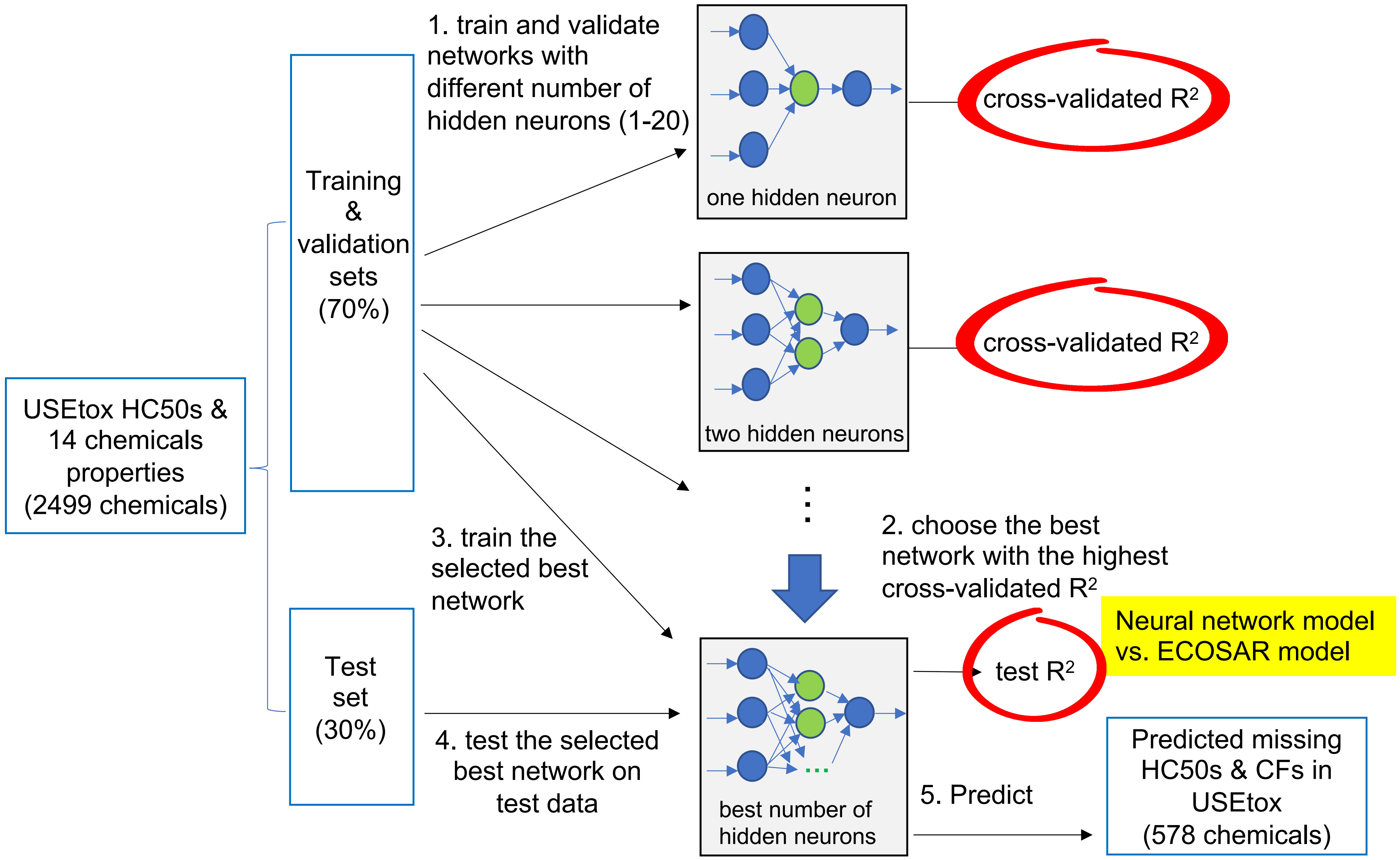
ECOSAR model test R^2: 0.194

Research question and goal

- Can we make use of the existing data to get a ***better estimation*** of HC50 (one kind of ecotoxicity) and ***avoid the time and cost*** of laboratory tests?

HC_{50} [kg·m^{-3}] is defined as the hazardous concentration of a chemical at which 50% of the freshwater species are exposed above their EC_{50}. The EC_{50} is the effective concentration at which 50% of a population displays an effect (e.g. mortality) in a laboratory test or a field test.

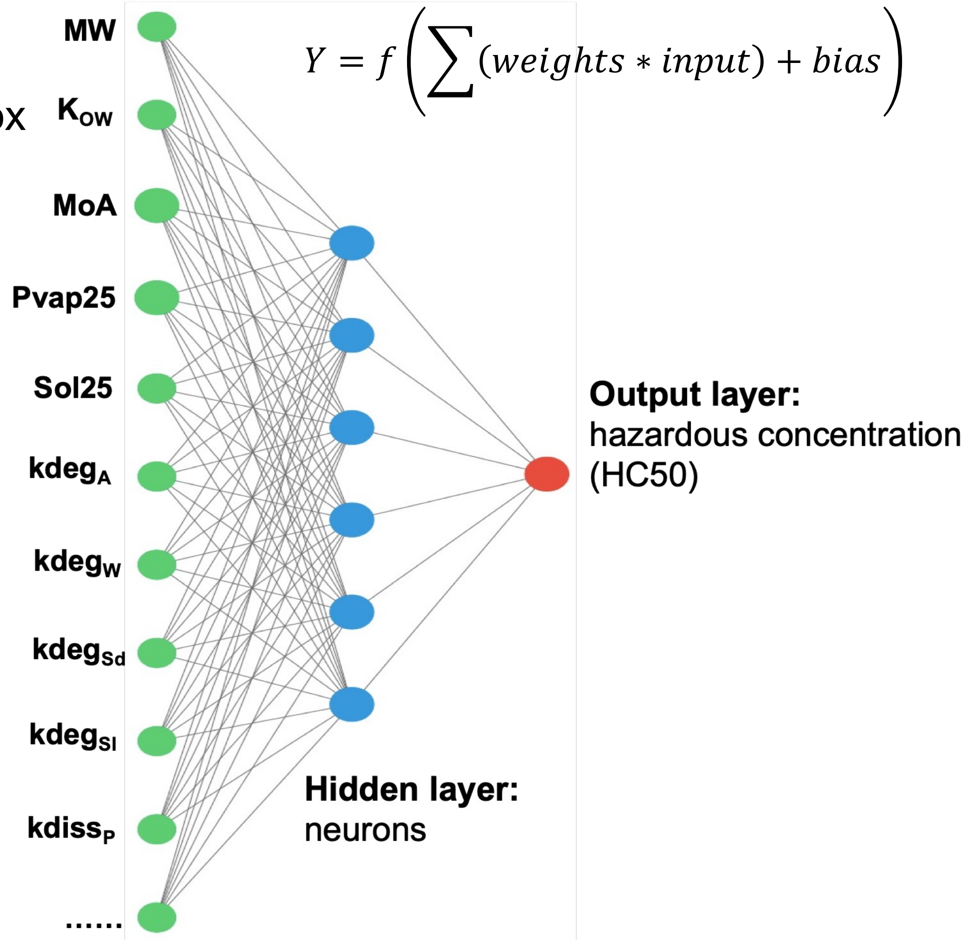
Steps of Building the Neural Network Model

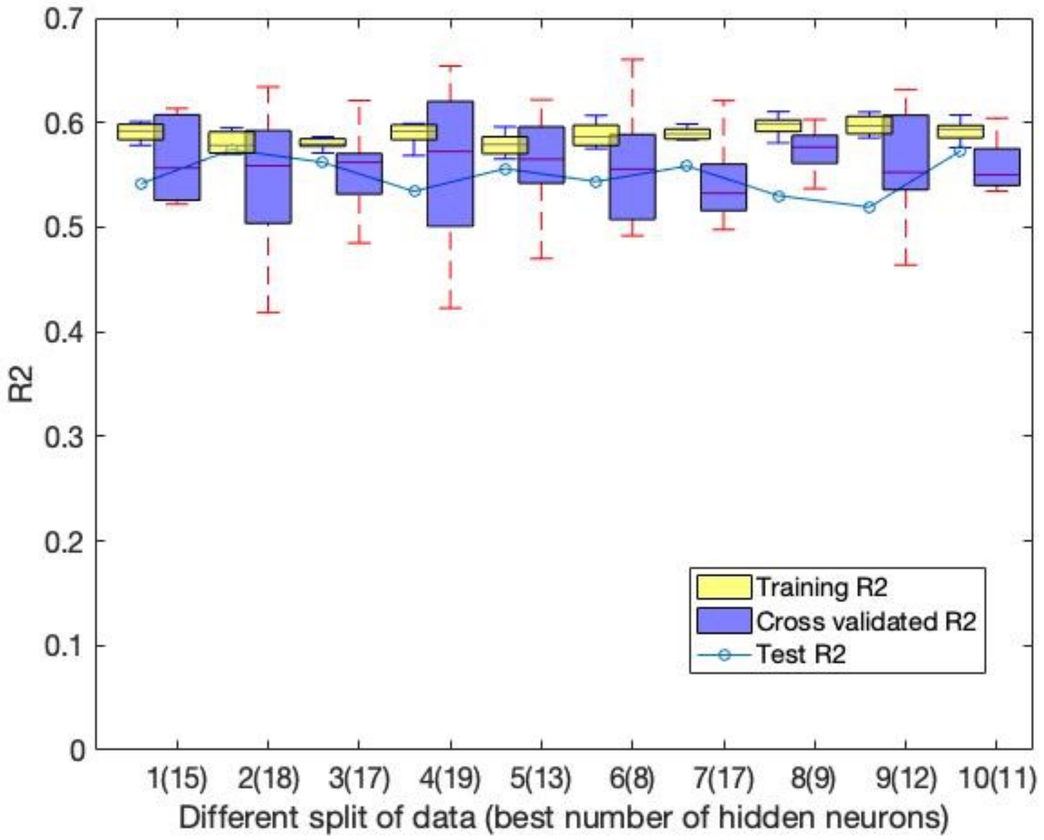


Results – ANN for predicting ecotoxicity

Input layer:
physicochemical properties in USEtox

Physical properties:
melting point
boiling point
water solubility
…
Chemical properties:
molecular weight
degradation rate
…





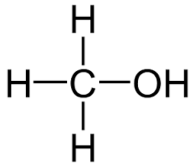
Neural network model avg. test R^2: 0.549

- Hou, P., Jolliet, O., Zhu, J., & Xu, M. (2020). Estimate ecotoxicity characterization factors for chemicals in life cycle assessment using machine learning models. *Environment international*, *135*, 105393.

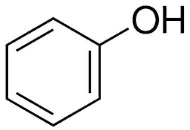
Improve the model with domain knowledge

- Classify chemicals into different mode of action (MoA) by Verharr scheme

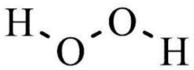
Class 1: Inert — Aliphatic alcohols, e.g., methanol



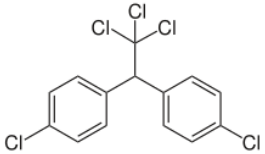
Class 2: Less inert — Phenols, anilines



Class 3: Reactive — Peroxide formers



Class 4: Specifically acting — Pesticides, herbicides

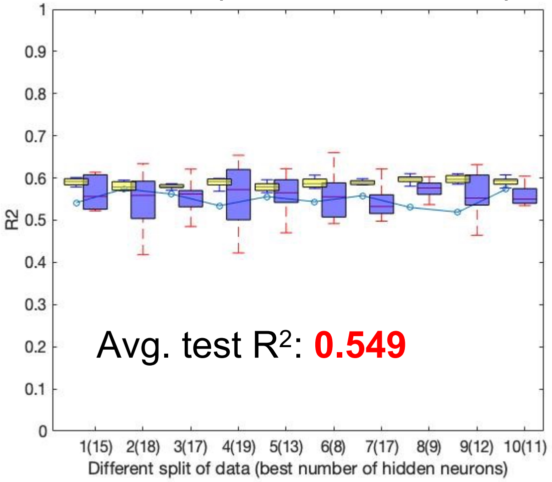


Class 5: Not classifiable

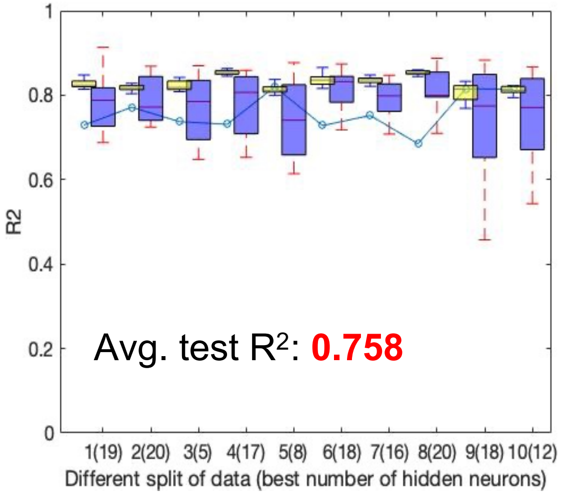
Toxicity increasing

Results – model performance by different MOA

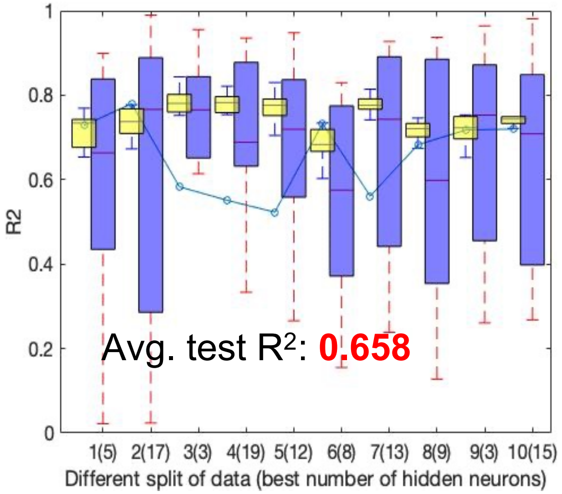
All data (2,308 samples)



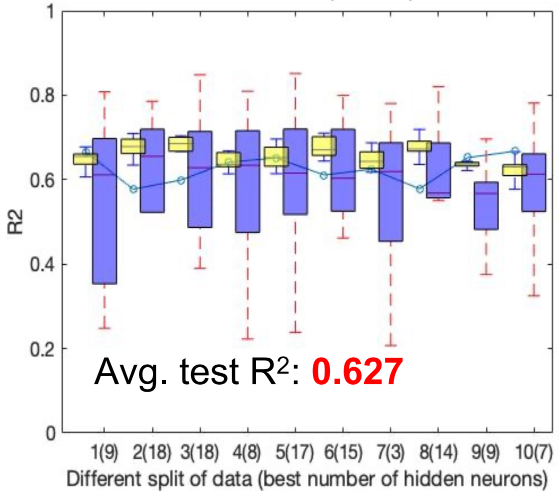
Class 1 (485)



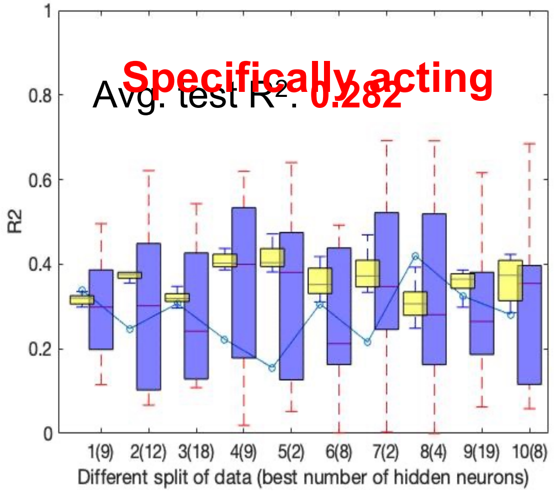
Class 2 (90)



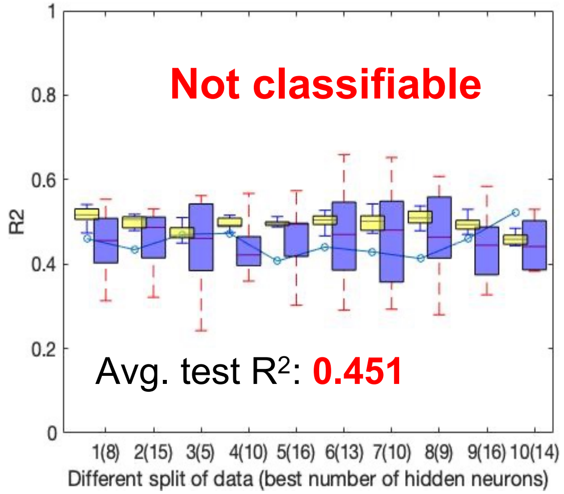
Class 3 (304)



Class 4 (195)



Class 5 (1,234)

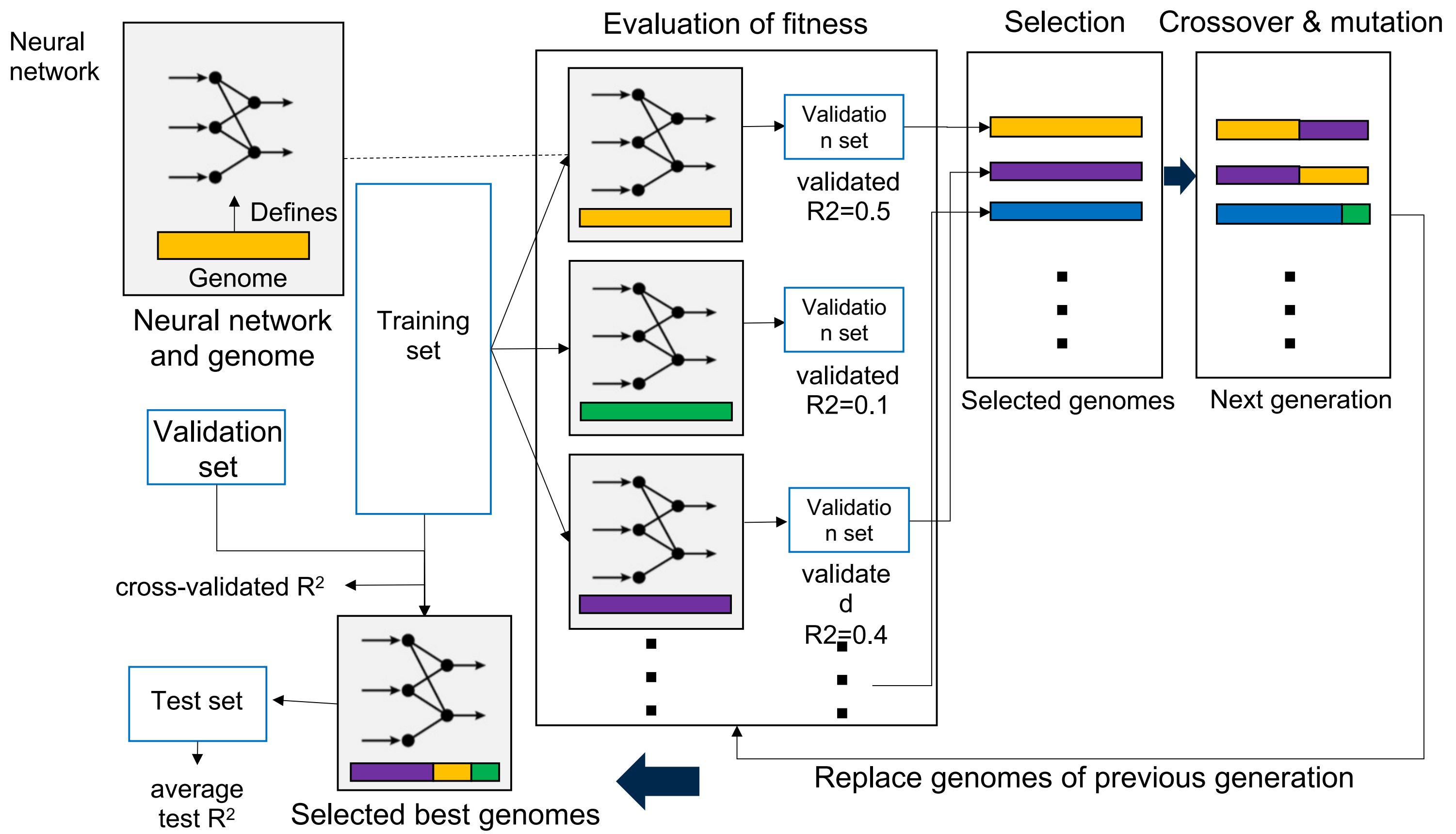


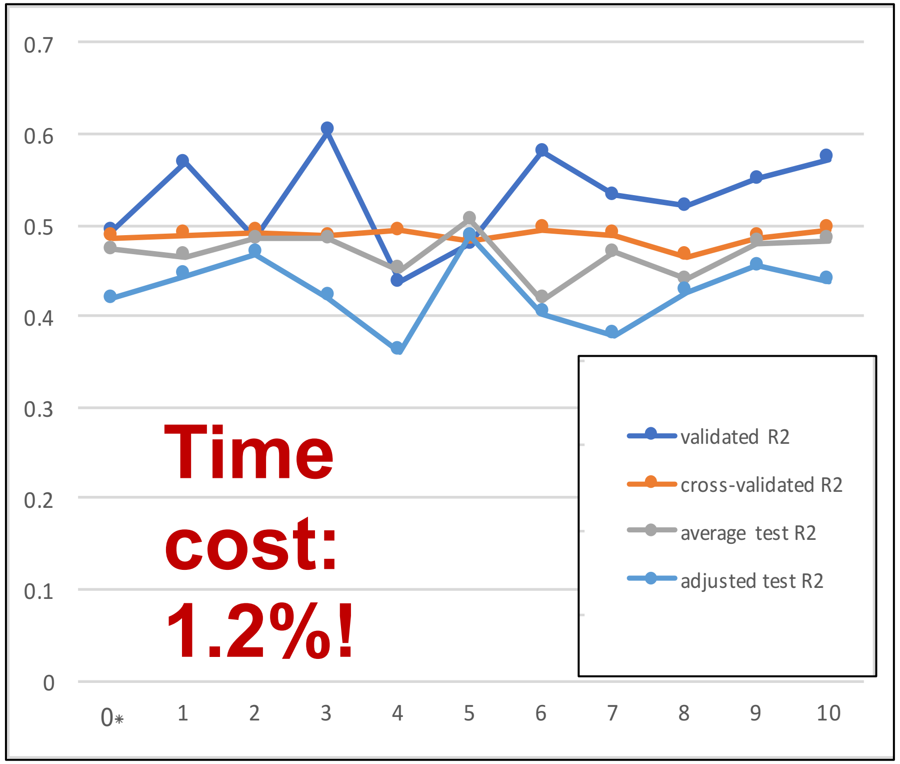
Hyperparameters in the neural network

Hyperparameters	Possible options
Number of hidden layers	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, …
Neurons per hidden layer	10, 20, 30, 40, 50, 60, 70, 80, 90, 100, …
Activation function	relu, elu, tanh, sigmoid, hard_sigmoid, softplus, linear
Network optimizer	rmsprop, adam, sgd, adagrad, adadelta, adamax, nadam

- How can we find the best neural network model among all parameter combinations?
  - Grid search: try all combinations
  - ***Genetic algorithm: a directed random search technique that simulates the natural selection and evolution process.***

Using genetic algorithm for optimization





- Genetic algorithm can find comparable performance networks with the brute force method.

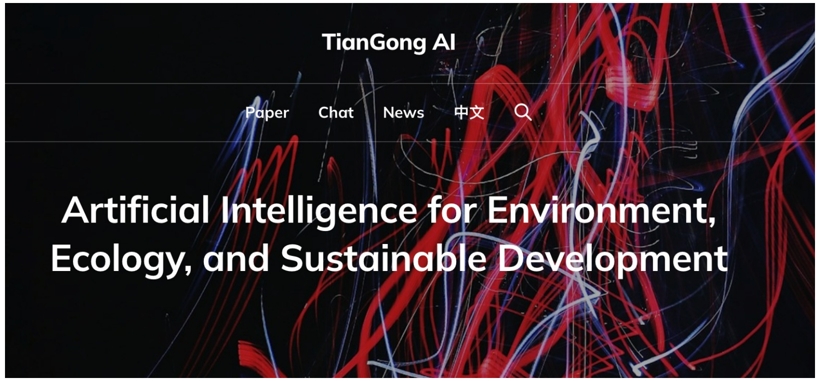
- Hou, P., Zhao, B., Jolliet, O., Zhu, J., Wang, P., & Xu, M. (2020). Rapid prediction of chemical ecotoxicity through genetic algorithm optimized neural network models. *ACS Sustainable Chemistry & Engineering*, *8*(32), 12168-12176.

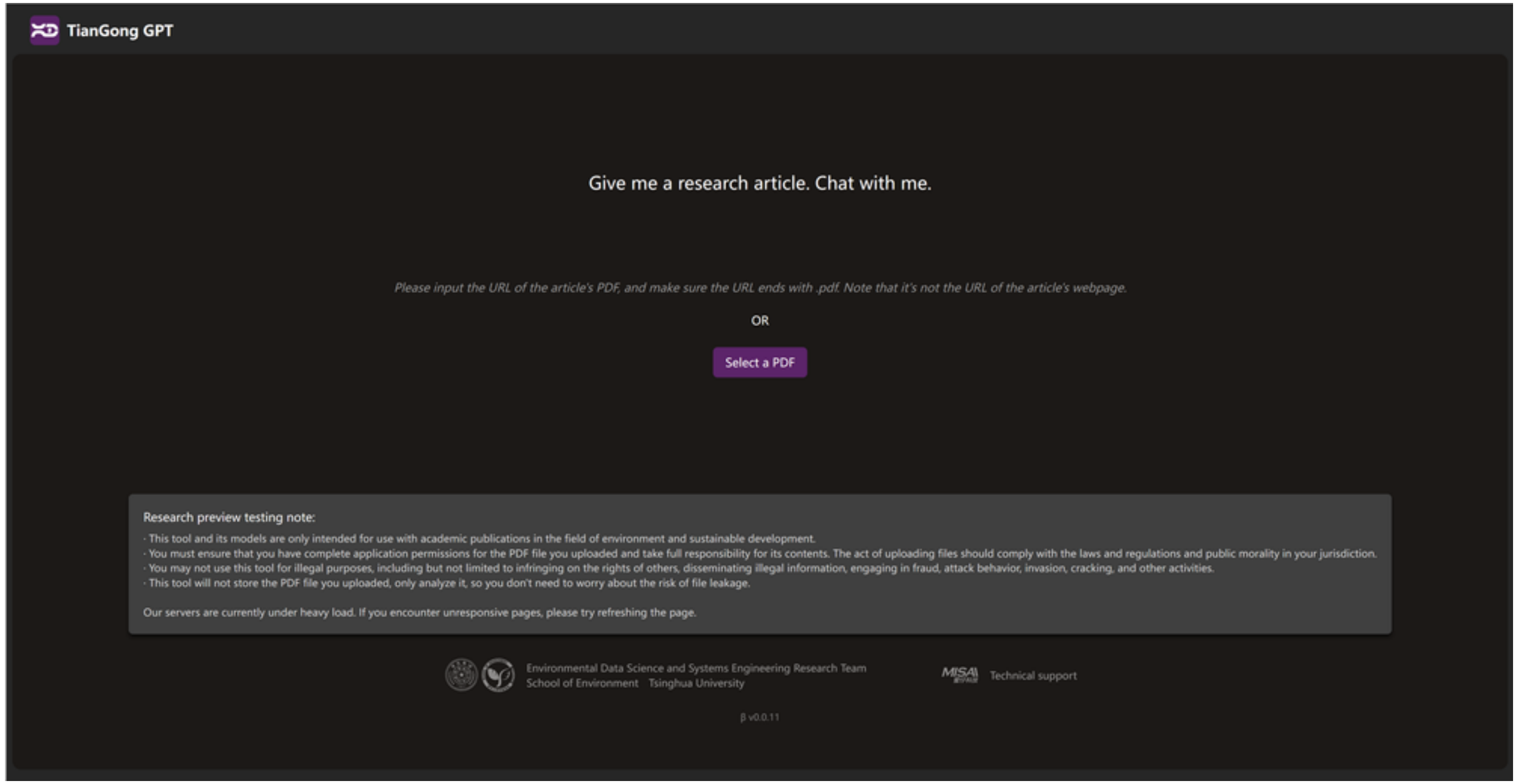
Summary

This study provides a machine learning model to estimate HC50 in USEtox to calculate characterization factors for chemicals based on their physical-chemical properties

1. Our model outperforms a traditional quantitative structure-activity relationship (QSAR) model (ECOSAR)
2. Use validated model to predict missing Cf_{eco} in USETox
3. Applied to a much border range of chemicals.

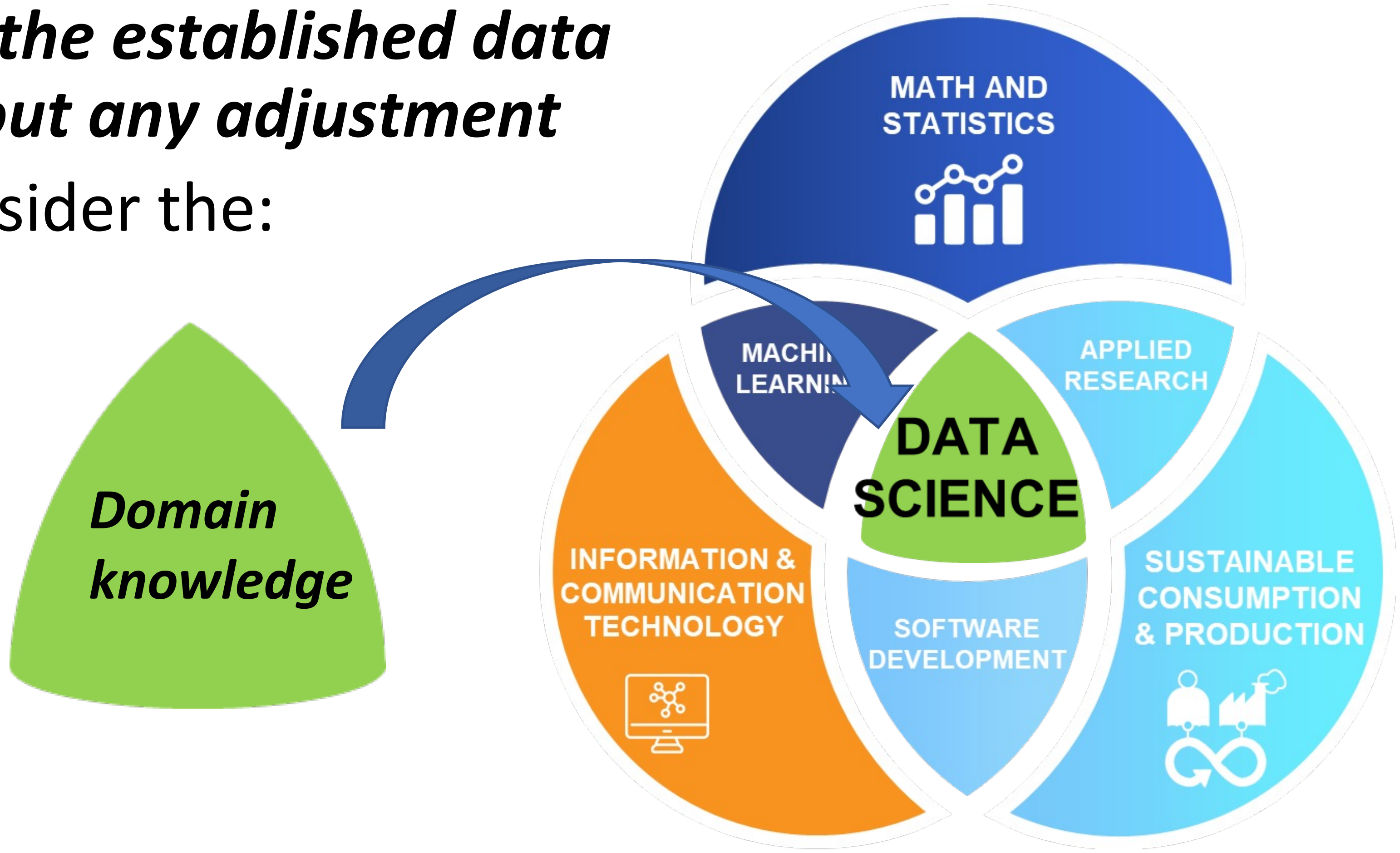
Other exploration





Concluding remarks

- Note for future study:
  - ***Cannot simply apply the established data science models without any adjustment***
  - Need to carefully consider the:
    - ***Objective***
    - ***Characteristics***
    - ***Particularity***
  - Choose properly***:***
    - ***Method***
    - ***Model structure***
    - ***Input features***
    - ***Response***



Publications

- Zhao, B., Jiang, J., Xu, M., & Tu Q. (2023). A Data-Centric Investigation on the Challenges of Similarity-Based Machine Learning Methods for Bridging Life Cycle Inventory Data Gap. Journal of Industrial Ecology. Under Review.

- Zhao, B., Shuai, C., Hou, P., Qu, S., & Xu, M. (2021). Estimation of unit process data for life cycle assessment using a decision tree-based approach. *Environmental Science & Technology*, *55*(12), 8439-8446.

- Hou, P., Zhao, B., Jolliet, O., Zhu, J., Wang, P., & Xu, M. (2020). Rapid prediction of chemical ecotoxicity through genetic algorithm optimized neural network models. *ACS Sustainable Chemistry & Engineering*, *8*(32), 12168-12176.

- Hou, P., Jolliet, O., Zhu, J., & Xu, M. (2020). Estimate ecotoxicity characterization factors for chemicals in life cycle assessment using machine learning models. *Environment international*, *135*, 105393.

- Hou, P., Cai, J., Qu, S., & Xu, M. (2018). Estimating missing unit process data in life cycle assessment using a similarity-based approach. *Environmental science & technology*, *52*(9), 5259-5267.

Thank you all for listening!

Bu Zhao, Ph.D.
Schmidt AI for Science Fellow
bz294@cornell.edu

School of Civil and Environmental Engineering
Cornell University, Ithaca