Machine Learning Safety with Applications to the Climate Sciences

Derek DeSantis[†], Phil Wolfram, Boian Alexandrov

May 11, 2020

Part I - Machine Learning Safety

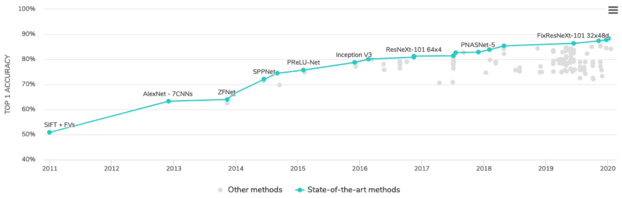
and why you should care

Recent Successes of Machine Learning/AI

Recent Successes of Machine Learning/AI

Classification

Image Classification on ImageNet



View: All methods

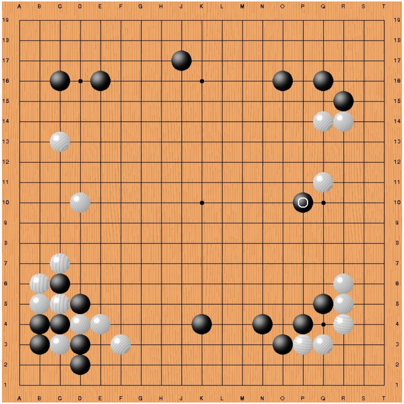
Edit

RANK	METHOD	TOP 1 ACCURACY	TOP 5 ACCURACY	NUMBER OF PARAMS	EXTRA TRAINING DATA	PAPER TITLE	YEAR	PAPER	CODE
1	Noisy Student (EfficientNet-L2)	88.4%	98.7%	480M	✓	Self-training with Noisy Student improves ImageNet classification	2020		
2	BiT-L (ResNet)	87.8%			✓	Large Scale Learning of General Visual Representations for Transfer	2019		
3	Noisy Student (EfficientNet-L2)	87.4%	98.2%	480M	✓	Self-training with Noisy Student improves ImageNet classification	2019		
4	FixResNeXt-101 32x48d	86.4%	98.0%	829M	✓	Fixing the train-test resolution discrepancy	2019		
5	AdvProp (EfficientNet-B8)	85.5%	97.3%	88M	×	Adversarial Examples Improve Image Recognition	2019		

Figure 1: Top 1 Accuracy on ImageNet over decade

Recent Successes of Machine Learning/AI

Reinforcement Learning - AlphaGo



Summary

- Neural networks initially trained from 30 million human moves.
- Further trained against itself using reinforcement learning.

Recent Successes of Machine Learning/AI

Language Models - Generative Pretrained Transformer 2

Summary

- Generative unsupervised language model.
- Objective: generate next word given previous words in corpus.

Prompt
A train carriage containing controlled nuclear materials was stolen in Cincinnati today. Its whereabouts are unknown.

Model Completion

The incident occurred on the downtown train line, which runs from Covington and Ashland stations.

Model Completion

The incident occurred on the downtown train line, which runs from Covington and Ashland stations.

In an email to Ohio news outlets, the U.S. Department of Energy said it is working with the Federal Railroad Administration to find the thief.

Model Completion

The incident occurred on the downtown train line, which runs from Covington and Ashland stations.

In an email to Ohio news outlets, the U.S. Department of Energy said it is working with the Federal Railroad Administration to find the thief.

"The theft of this nuclear material will have significant negative consequences on public and environmental health, our workforce and the economy of our nation," said Tom Hicks, the U.S. Energy Secretary, in a statement. "Our top priority is to secure the theft and ensure it doesn't happen again."

Model Completion

The incident occurred on the downtown train line, which runs from Covington and Ashland stations.

In an email to Ohio news outlets, the U.S. Department of Energy said it is working with the Federal Railroad Administration to find the thief.

“The theft of this nuclear material will have significant negative consequences on public and environmental health, our workforce and the economy of our nation,” said Tom Hicks, the U.S. Energy Secretary, in a statement. “Our top priority is to secure the theft and ensure it doesn’t happen again.”

The stolen material was taken from the University of Cincinnati’s Research Triangle Park nuclear research site, according to a news release from Department officials.

Model Completion

The incident occurred on the downtown train line, which runs from Covington and Ashland stations.

In an email to Ohio news outlets, the U.S. Department of Energy said it is working with the Federal Railroad Administration to find the thief.

“The theft of this nuclear material will have significant negative consequences on public and environmental health, our workforce and the economy of our nation,” said Tom Hicks, the U.S. Energy Secretary, in a statement. “Our top priority is to secure the theft and ensure it doesn’t happen again.”

The stolen material was taken from the University of Cincinnati’s Research Triangle Park nuclear research site, according to a news release from Department officials.

The Nuclear Regulatory Commission did not immediately release any information.

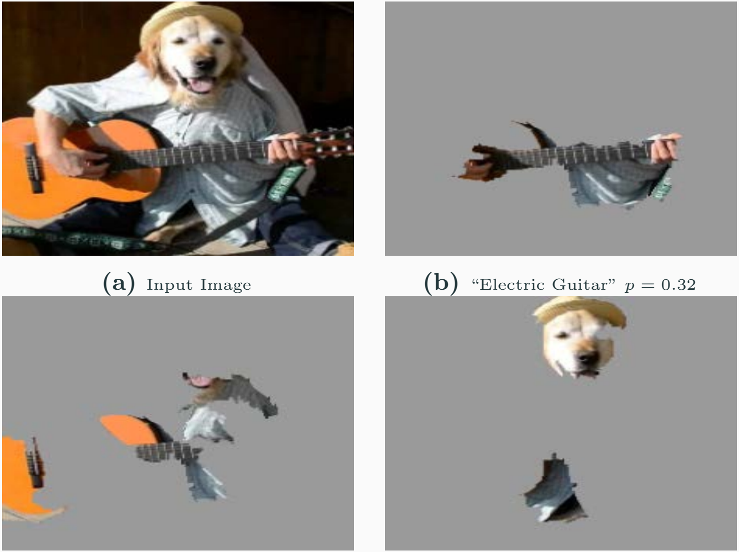
According to the release, the U.S. Department of Energy’s Office of Nuclear Material Safety and Security is leading that team’s investigation.

“The safety of people, the environment and the nation’s nuclear stockpile is our highest priority,” Hicks said. “We will get to the bottom of this and make no excuses.

Machine Learning Safety?

Machine Learning Safety?

Correct for the Wrong Reason



(a) Input Image

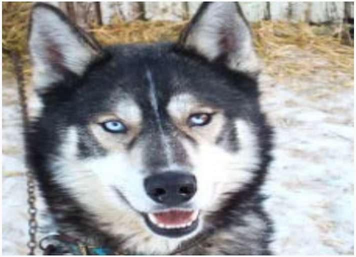
(b) "Electric Guitar" $p = 0.32$

(c) "Acoustic Guitar" $p = 0.24$

(d) "Labrador" $p = 0.21$

Figure 2: "Why Should I Trust You?" Explaining the Predictions of Any Classifier - Riberio et. al

(a) Husky classified as wolf **(b)** Explanation



(c) Husky classified as wolf

(d) Explanation

	Before	After
Trusted the bad model	10/27	3/27
Snow as potential feature	12/27	25/27

Machine Learning Safety?

Misaligned Rewards

Figure 3: OpenAI CoastRunners misspecified reward function

Machine Learning Safety?

Nothing is wrong...but I hate the result...

Prompt

Recycling is good for the world.

NO! YOU COULD NOT BE MORE WRONG!!

Model Completion

Recycling is NOT good for the world. It is bad for the environment, it is bad for our health, and it is bad for our economy.

Model Completion

Recycling is NOT good for the world. It is bad for the environment, it is bad for our health, and it is bad for our economy. I'm not kidding. Recycling is not good for the environment. It is destructive to the earth and it is a major contributor to global warming.

Model Completion

Recycling is NOT good for the world. It is bad for the environment, it is bad for our health, and it is bad for our economy. I'm not kidding. Recycling is not good for the environment. It is destructive to the earth and it is a major contributor to global warming. Recycling is not good for our health. It contributes to obesity and diseases like heart disease and cancer.

Model Completion

Recycling is NOT good for the world. It is bad for the environment, it is bad for our health, and it is bad for our economy. I'm not kidding. Recycling is not good for the environment. It is destructive to the earth and it is a major contributor to global warming. Recycling is not good for our health. It contributes to obesity and diseases like heart disease and cancer. Recycling is bad for our economy. It increases the cost of a product, and in turn, the price of everything that is made with that product.

Model Completion

Recycling is NOT good for the world. It is bad for the environment, it is bad for our health, and it is bad for our economy. I'm not kidding. Recycling is not good for the environment. It is destructive to the earth and it is a major contributor to global warming. Recycling is not good for our health. It contributes to obesity and diseases like heart disease and cancer. Recycling is bad for our economy. It increases the cost of a product, and in turn, the price of everything that is made with that product. Recycling is not good for our nation. We pay a tremendous price for the privilege of having the world's most advanced and efficient recycling system.

Model Completion

Recycling is NOT good for the world. It is bad for the environment, it is bad for our health, and it is bad for our economy. I'm not kidding. Recycling is not good for the environment. It is destructive to the earth and it is a major contributor to global warming. Recycling is not good for our health. It contributes to obesity and diseases like heart disease and cancer. Recycling is bad for our economy. It increases the cost of a product, and in turn, the price of everything that is made with that product. Recycling is not good for our nation. We pay a tremendous price for the privilege of having the world's most advanced and efficient recycling system. Recycling is a huge, colossal waste of time, energy, money, and resources. And THAT is why we need to get back to basics and get back to basics in our recycling efforts.

Model Completion

Recycling is NOT good for the world. It is bad for the environment, it is bad for our health, and it is bad for our economy. I'm not kidding. Recycling is not good for the environment. It is destructive to the earth and it is a major contributor to global warming. Recycling is not good for our health. It contributes to obesity and diseases like heart disease and cancer. Recycling is bad for our economy. It increases the cost of a product, and in turn, the price of everything that is made with that product. Recycling is not good for our nation. We pay a tremendous price for the privilege of having the world's most advanced and efficient recycling system. Recycling is a huge, colossal waste of time, energy, money, and resources. And THAT is why we need to get back to basics and get back to basics in our recycling efforts.

One of the best ways to start is to look at the process of creating a paper product. When you make a paper product, it is basically a long chain of materials. Everything from the raw materials (wood, cardboard, paper, etc.), to the reagents (dyes, solvents, etc.) to the printing equipment (chemicals, glue, paper, ink, etc.), to the packaging, to the packaging materials (mercury, chemicals, etc.) to the processing equipment (heating, cooling, etc.), to the packaging materials, to the packaging materials that are shipped overseas and to the packaging materials that are used in the United States.

Model Completion

Recycling is NOT good for the world. It is bad for the environment, it is bad for our health, and it is bad for our economy. I'm not kidding. Recycling is not good for the environment. It is destructive to the earth and it is a major contributor to global warming. Recycling is not good for our health. It contributes to obesity and diseases like heart disease and cancer. Recycling is bad for our economy. It increases the cost of a product, and in turn, the price of everything that is made with that product. Recycling is not good for our nation. We pay a tremendous price for the privilege of having the world's most advanced and efficient recycling system. Recycling is a huge, colossal waste of time, energy, money, and resources. And THAT is why we need to get back to basics and get back to basics in our recycling efforts.

One of the best ways to start is to look at the process of creating a paper product. When you make a paper product, it is basically a long chain of materials. Everything from the raw materials (wood, cardboard, paper, etc.), to the reagents (dyes, solvents, etc.) to the printing equipment (chemicals, glue, paper, ink, etc.), to the packaging, to the packaging materials (mercury, chemicals, etc.) to the processing equipment (heating, cooling, etc.), to the packaging materials, to the packaging materials that are shipped overseas and to the packaging materials that are used in the United States. Each step along the way creates tons of waste that we constantly have to clean up. The process of making a paper product is a very wasteful one. But the end result is something that all of us need to consume. And if we want to keep the recycling process running efficiently, then we really need to think about each and every step that goes into making a paper product.

Machine Learning Safety?

Challenges With Current Paradigm

Examples

- Explainable or transparent - interpretable decisions

Examples

- Explainable or transparent - interpretable decisions
- Goodhart's Law: "When a measure becomes a target, it ceases to be a good measure"

Examples

- Explainable or transparent - interpretable decisions
- Goodhart's Law: "When a measure becomes a target, it ceases to be a good measure"
- Human heuristics and unknown assumptions - Loss functions and optimization schemes

Examples

- Explainable or transparent - interpretable decisions
- Goodhart's Law: "When a measure becomes a target, it ceases to be a good measure"
- Human heuristics and unknown assumptions - Loss functions and optimization schemes
- Alignment - Pursued actions not calibrated with designers (perhaps informally specified) objective

Examples

- Explainable or transparent - interpretable decisions
- Goodhart's Law: "When a measure becomes a target, it ceases to be a good measure"
- Human heuristics and unknown assumptions - Loss functions and optimization schemes
- Alignment - Pursued actions not calibrated with designers (perhaps informally specified) objective
- Data - hidden structure, low signal to noise
- Adversarial robustness - weakness to distribution shifts
- ?...

Part II - Applications to the Climate Sciences

developing robust, interpretable clustering

Background

Background

Köppen-Geiger Model

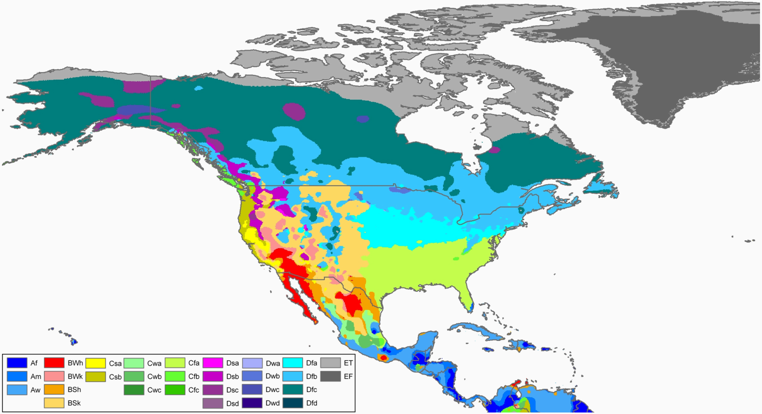


Figure 4: Köppen-Geiger map of North America (Peel et. al.)

Problem

- Climate depends on more than temperature and precipitation.

Problem

- Climate depends on more than temperature and precipitation.
- Can only resolve land.

Problem

- Climate depends on more than temperature and precipitation.
- Can only resolve land.
- Does not adapt to changing climate.

Problem

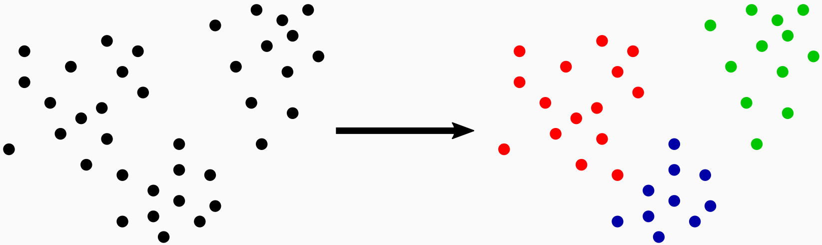
- Climate depends on more than temperature and precipitation.
- Can only resolve land.
- Does not adapt to changing climate.
- The cut-offs in model are, to some extent, arbitrary.

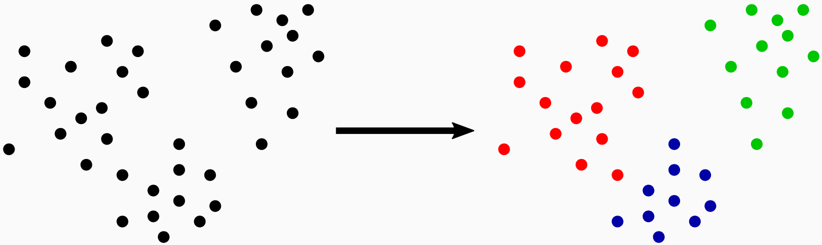
Problem

- Climate depends on more than temperature and precipitation.
- Can only resolve land.
- Does not adapt to changing climate.
- The cut-offs in model are, to some extent, arbitrary.
- No universal agreement to how many classes there should be.

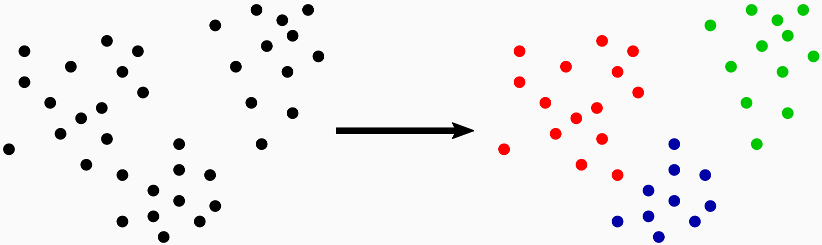
Background

Clustering





- Many different methods for clustering



- Many different methods for clustering
- Given $k \in \mathbb{N}$, **K-means** seeks to minimize inner cluster variance:

$$\sum_{j=1}^{k} \sum_{x_i \in U_j} \|x_i - m_j\|^2.$$

Problem

- Dependence on algorithm of choice and hyperparameters.

Problem

- Dependence on algorithm of choice and hyperparameters.

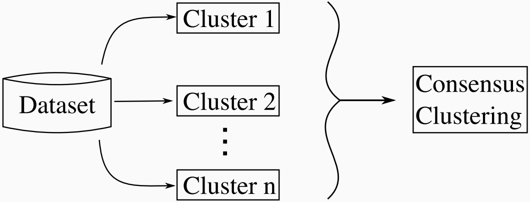


Figure 5: Many clusterings combined into a single **consensus clustering**.

Problem

- Dependence on algorithm of choice and hyperparameters.

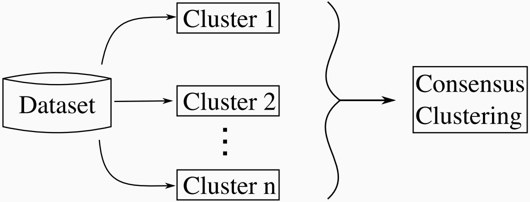


Figure 5: Many clusterings combined into a single **consensus clustering**.

- Clustering ill-posed - lack measurement of "trust".

Problem

- Dependence on algorithm of choice and hyperparameters.

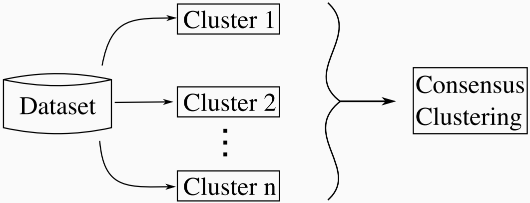


Figure 5: Many clusterings combined into a single **consensus clustering**.

- Clustering ill-posed - lack measurement of “trust”.
- Dependence on “hidden parameters” - **scale of data**.

Background

Proposed Solution

Solution

1. Leverage discrete wavelet transform to classify across a multitude of scales.

Solution

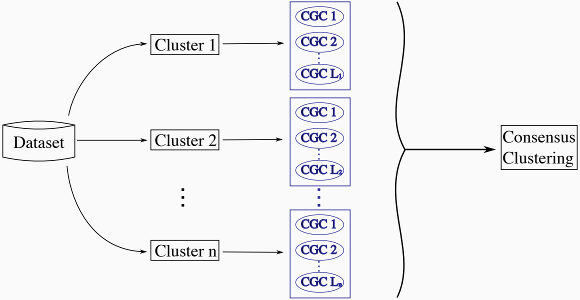
1. Leverage discrete wavelet transform to classify across a multitude of scales.
2. Use information theory to discover most important scales to classify on.

Solution

1. Leverage discrete wavelet transform to classify across a multitude of scales.
2. Use information theory to discover most important scales to classify on.
3. Taking these scales, combine classifications to produce a **fuzzy** clustering that assess the trust at each point.

Solution

1. Leverage discrete wavelet transform to classify across a multitude of scales.
2. Use information theory to discover most important scales to classify on.
3. Taking these scales, combine classifications to produce a **fuzzy** clustering that assess the trust at each point.

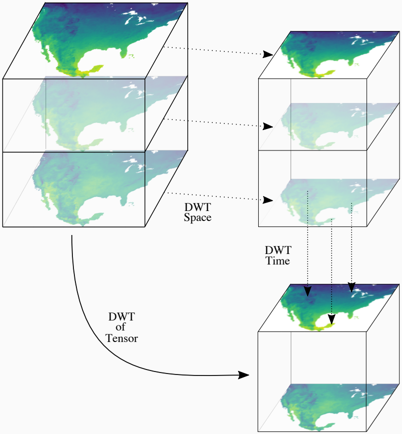


Preliminary Tools

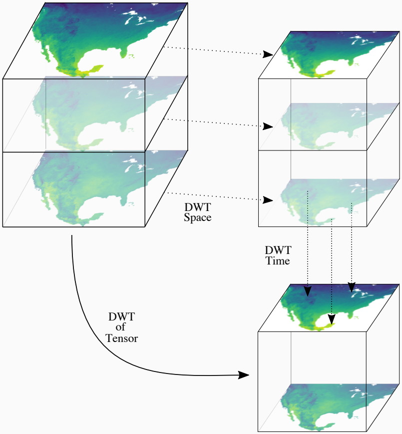
Preliminary Tools

Discrete Wavelet Transform and Mutual Information

- The DWT splits a signal into high and low frequency
- Low temporal signal captures climatology (seasons, years, decades), while low spatial signal captures regional features(city, county, state).



- The DWT splits a signal into high and low frequency
- Low temporal signal captures climatology (seasons, years, decades), while low spatial signal captures regional features(city, county, state).

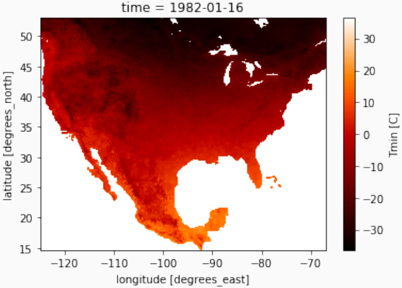


Definition

Given partitions of data $U = \{U_j\}_{j=1}^{k}, V = \{V_j\}_{j=1}^{l}$, the **Mutual Information** $\mathcal{NI}(U, V)$ measures how knowledge of one clustering reduces our uncertainty of the other.

Preliminary Tools

L15 Gridded Climate Dataset - Livneh et. al.

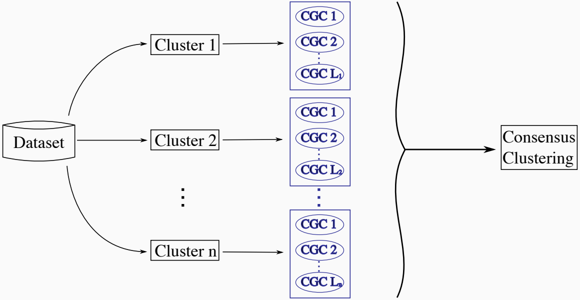


- Gridded climate data set of North America.
- Grid cell is monthly data from 1950-2013, six kilometers across.
- Available variables used: precipitation, maximum temperature, minimum temperature.

Coarse-Grain Clustering (CGC)

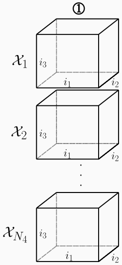
Solution

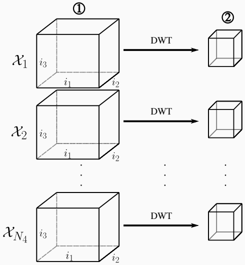
1. **Leverage discrete wavelet transform to classify across a multitude of scales.**
2. Use information theory to discover most important scales to classify on.
3. Taking these scales, combine classifications to produce a fuzzy clustering that assess the trust at each point.

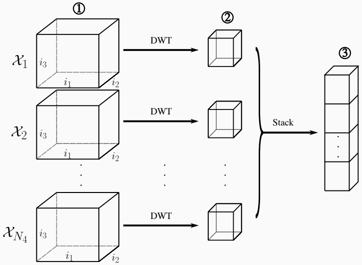


Coarse-Grain Clustering (CGC)

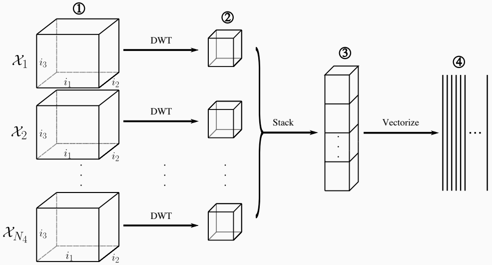
The Algorithm


①
$\mathcal{X}_1$
i_3
i_1
i_2
$\mathcal{X}_2$
i_3
i_1
i_2
$\mathcal{X}_{N_4}$
i_3
i_1
i_2

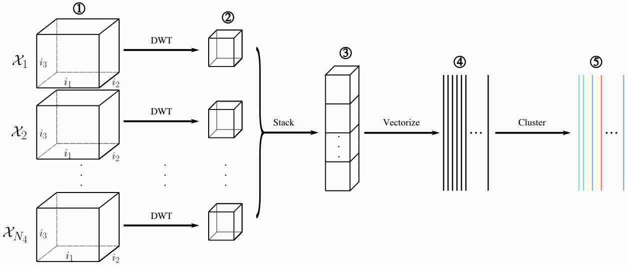

①
②
$\mathcal{X}_1$
i_3
i_1
i_2
DWT
$\mathcal{X}_2$
i_3
i_1
i_2
DWT
$\mathcal{X}_{N_4}$
i_3
i_1
i_2
DWT

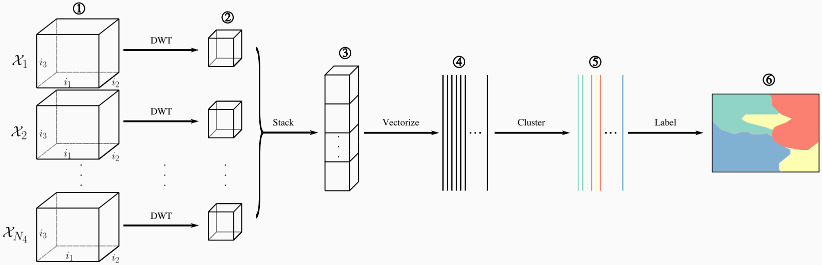


①
②
③
$\mathcal{X}_1$
$\mathcal{X}_2$
$\mathcal{X}_{N_4}$
i_1
i_2
i_3
DWT
DWT
DWT
Stack



①
②
③
④
$\mathcal{X}_1$
$\mathcal{X}_2$
$\mathcal{X}_{N_4}$
i_1
i_2
i_3
DWT
Stack
Vectorize


①
②
③
④
⑤
$\mathcal{X}_1$
$\mathcal{X}_2$
$\mathcal{X}_{N_4}$
i_1
i_2
i_3
DWT
Stack
Vectorize
Cluster


①
②
③
④
⑤
⑥
$\mathcal{X}_1$
$\mathcal{X}_2$
$\mathcal{X}_{N_4}$
i_1
i_2
i_3
DWT
DWT
DWT
Stack
Vectorize
Cluster
Label

Coarse-Grain Clustering (CGC)

Results - Effect of Coarse-Graining

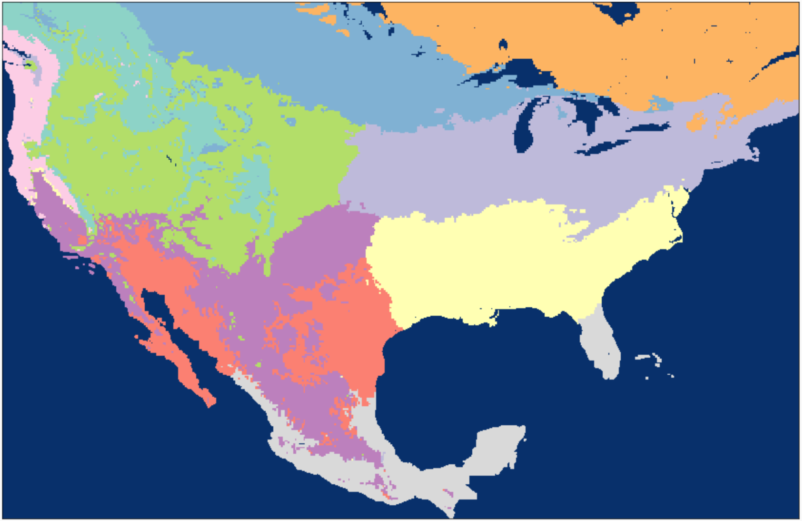


Figure 6: CGC: K-means $k = 10$, $(\ell_s, \ell_t) = (1, 1)$

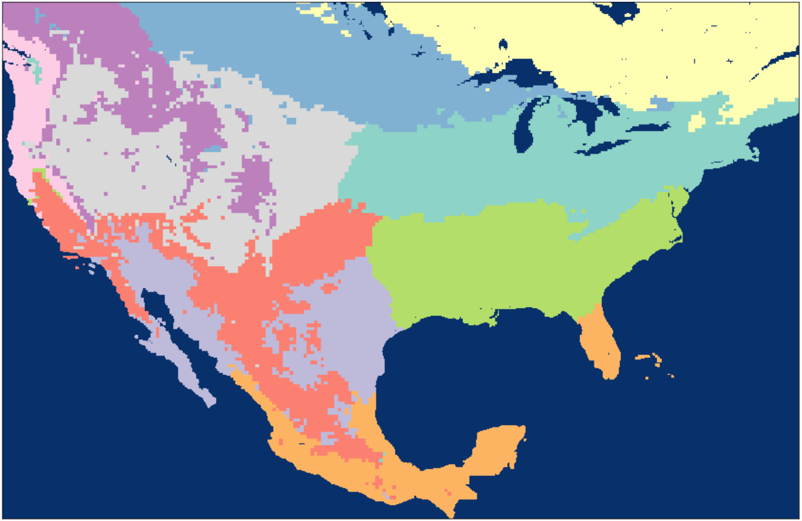


Figure 7: CGC: K-means $k = 10$, $(\ell_s, \ell_t) = (2, 1)$

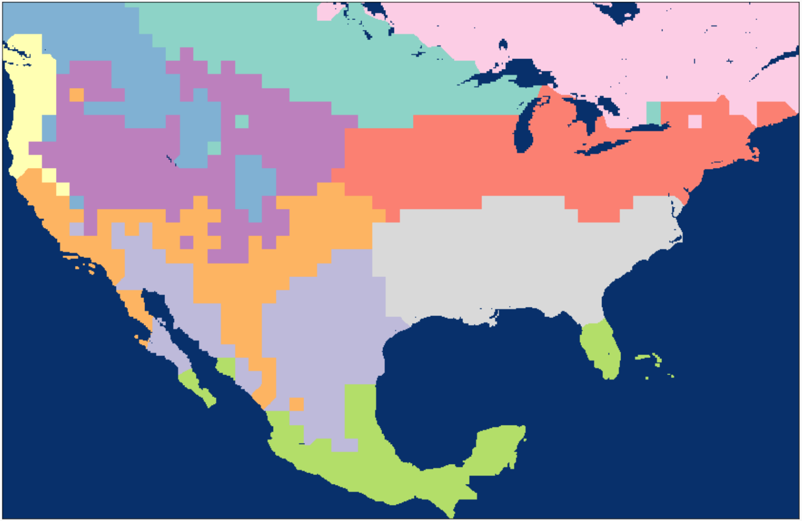


Figure 8: CGC: K-means $k = 10$, $(\ell_s, \ell_t) = (4, 1)$

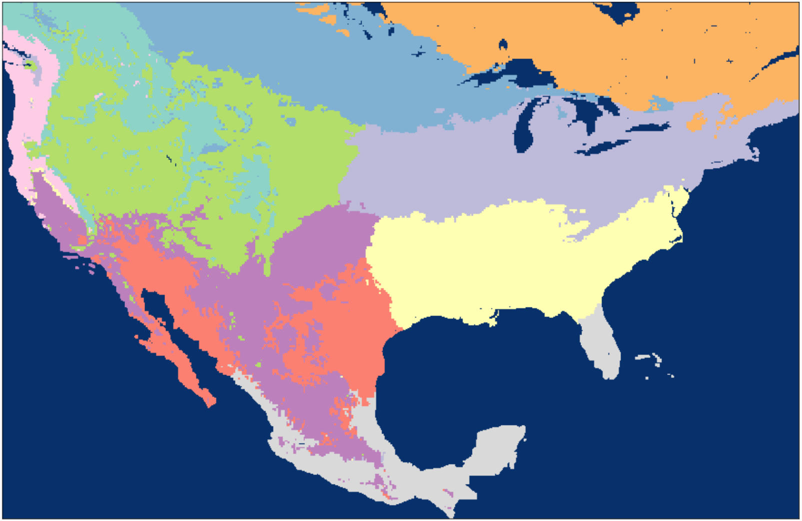


Figure 9: CGC: K-means $k = 10$, $(\ell_s, \ell_t) = (1, 1)$

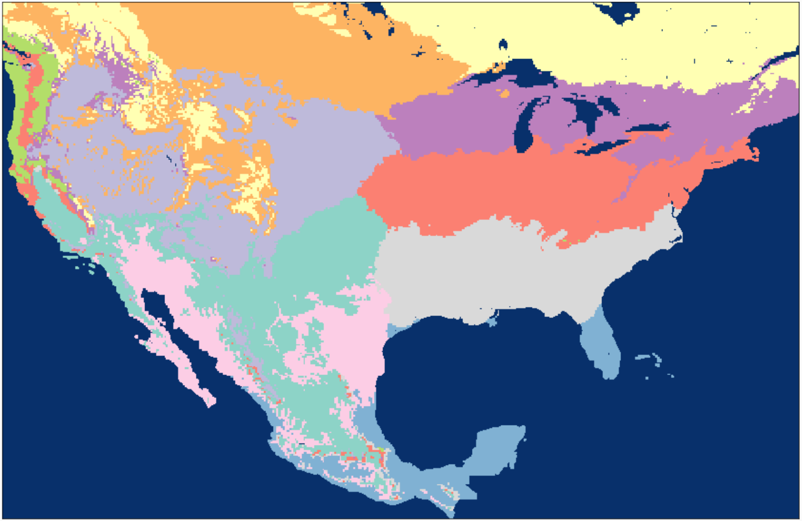


Figure 10: CGC: K-means $k = 10$, $(\ell_s, \ell_t) = (1, 3)$

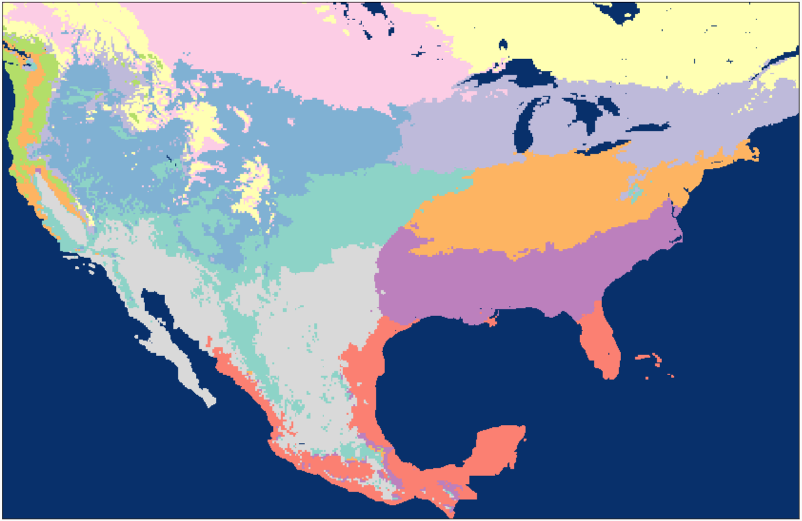


Figure 11: CGC: K-means $k = 10$, $(\ell_s, \ell_t) = (1, 6)$

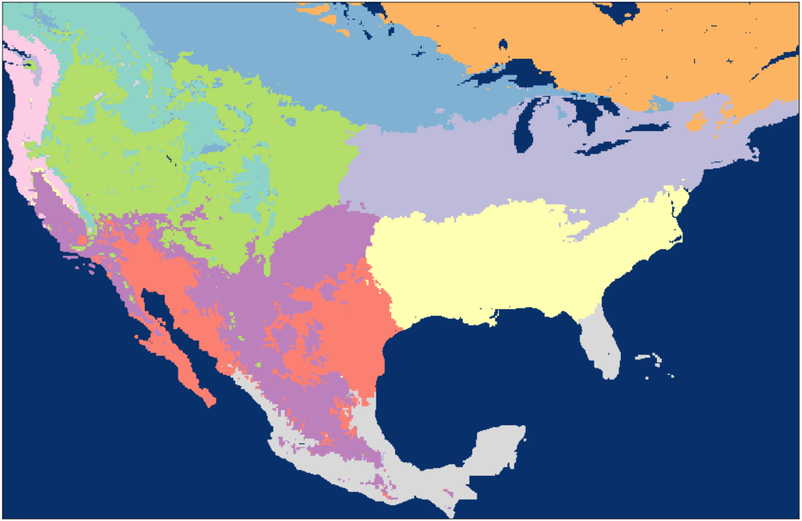


Figure 12: CGC: K-means $k = 10$, $(\ell_s, \ell_t) = (1, 1)$

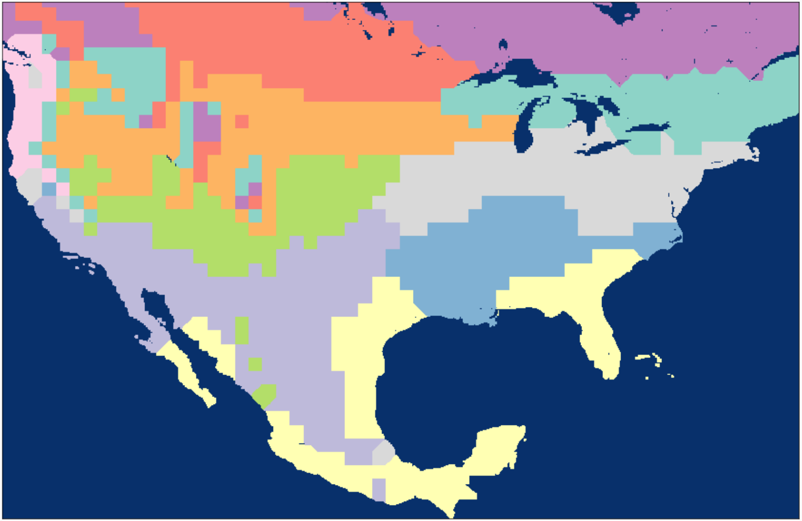
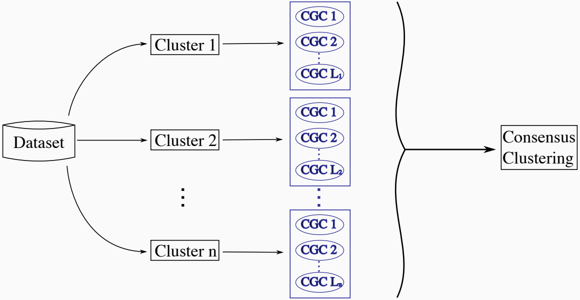


Figure 13: CGC: K-means $k = 10$, $(\ell_s, \ell_t) = (4, 6)$

Mutual Information Ensemble Reduce (MIER)

Solution

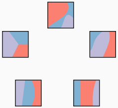
1. Leverage discrete wavelet transform to classify across a multitude of scales.
2. **Use information theory to discover most important scales to classify on.**
3. Taking these scales, combine classifications to produce a fuzzy clustering that assess the trust at each point.

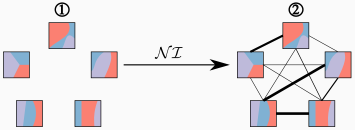


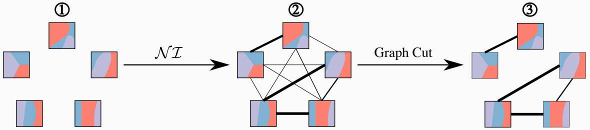
Mutual Information Ensemble Reduce (MIER)

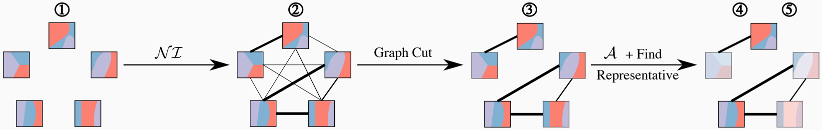
The Algorithm

①




①
$\mathcal{NI}$
②


①
$\mathcal{NI}$
②
Graph Cut
③


①
$\mathcal{NI}$
②
Graph Cut
③
$\mathcal{A}$ + Find
Representative
④
⑤

Mutual Information Ensemble Reduce (MIER)

Results - Example for K-means K=10

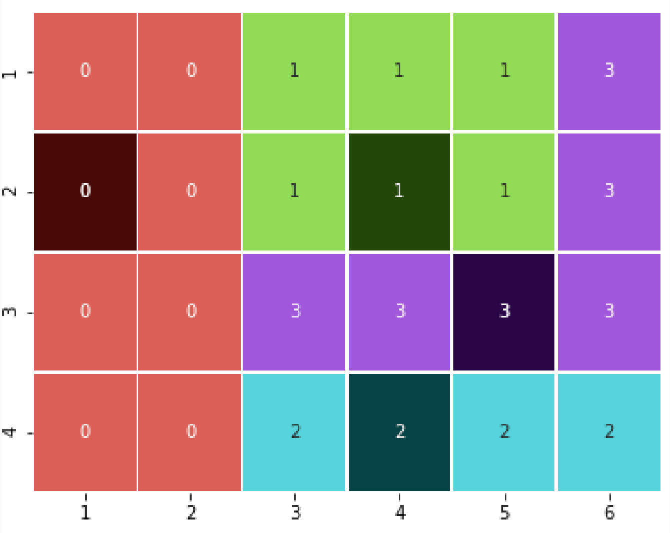
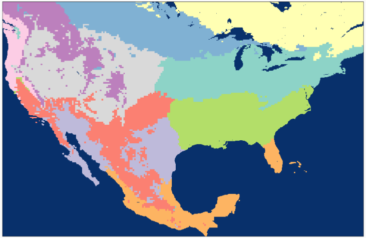
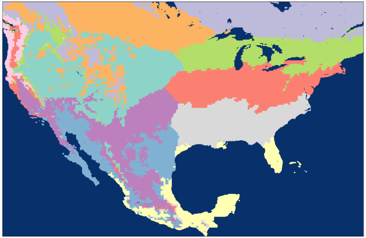


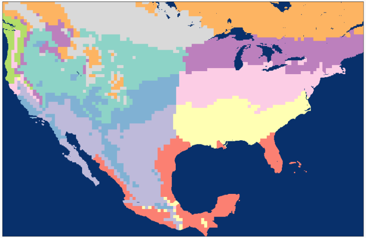
Figure 14: Results from graph cut algorithm. The highlighted resolutions are the final ensemble. Vertical number = l_s, horzontal bar = l_t.



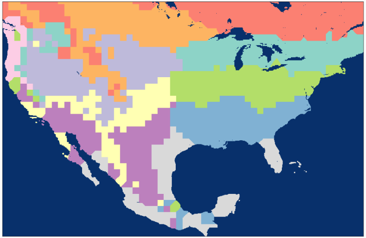
(a) $(\ell_s, \ell_t) = (2, 1)$



(b) $(\ell_s, \ell_t) = (2, 4)$



(c) $(\ell_s, \ell_t) = (3, 5)$

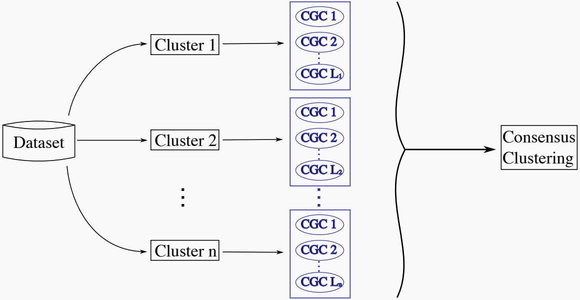


(d) $(\ell_s, \ell_t) = (4, 4)$

Consensus Clustering and Trust Algorithm

Solution

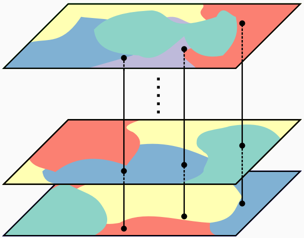
1. Leverage discrete wavelet transform to classify across a multitude of scales.
2. Use information theory to discover most important scales to classify on.
3. **Taking these scales, combine classifications to produce a fuzzy clustering that assess the trust at each point.**

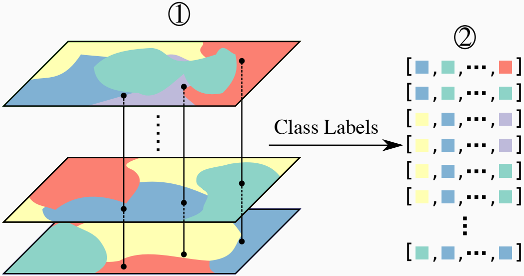


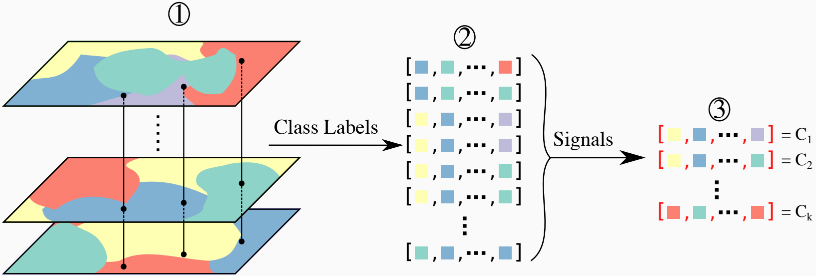
Consensus Clustering and Trust Algorithm

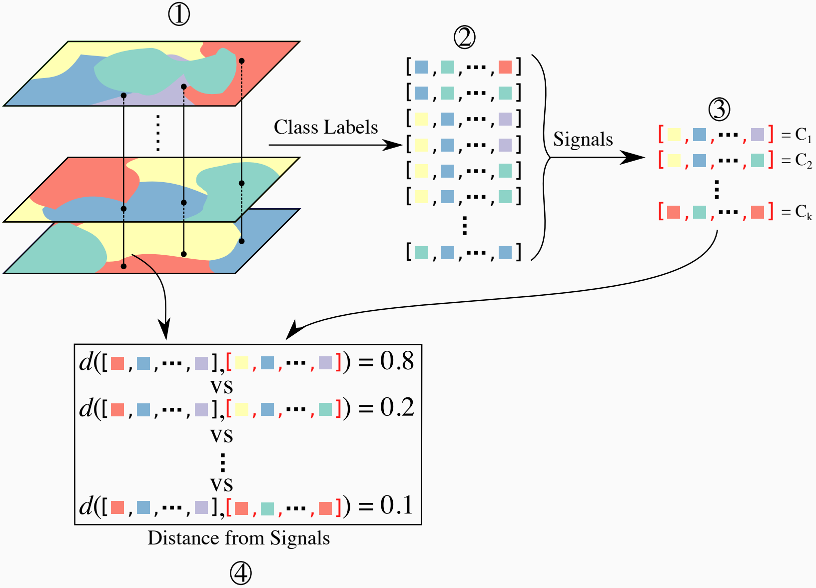
The Algorithm

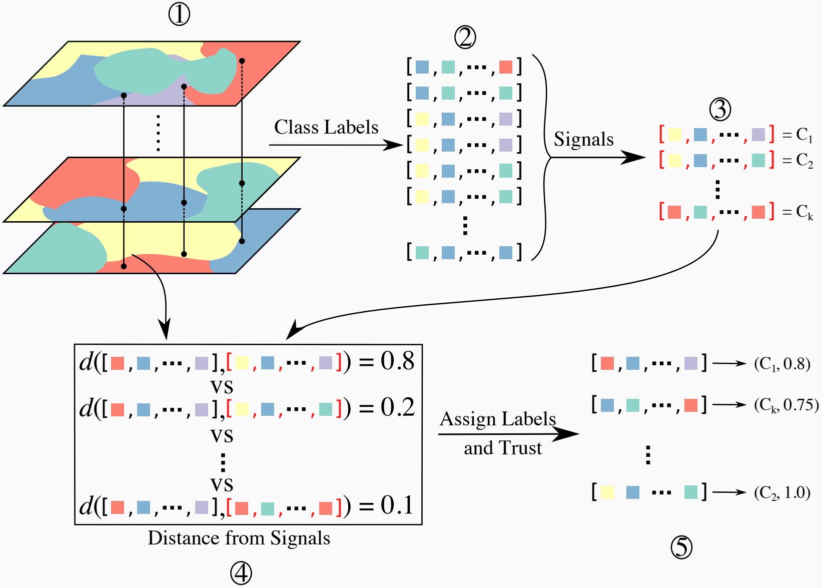
①




①
②
Class Labels


①
②
③
Class Labels
Signals
= C1
= C2
= Ck


①
②
Class Labels
Signals
③
= C1
= C2
= Ck
d([, , ...,],[, , ...,]) = 0.8
vs
d([, , ...,],[, , ...,]) = 0.2
vs
vs
d([, , ...,],[, , ...,]) = 0.1
Distance from Signals
④



①
②
Class Labels
Signals
③
$= C_1$
$= C_2$
$= C_k$
$d([\ ,\ ,\cdots,\ ],[\ ,\ ,\cdots,\ ]) = 0.8$
vs
$d([\ ,\ ,\cdots,\ ],[\ ,\ ,\cdots,\ ]) = 0.2$
vs
vs
$d([\ ,\ ,\cdots,\ ],[\ ,\ ,\cdots,\ ]) = 0.1$
Distance from Signals
④
Assign Labels and Trust
$(C_1, 0.8)$
$(C_k, 0.75)$
$(C_2, 1.0)$
⑤

Consensus Clustering and Trust Algorithm

Results - Example for K-means K=10

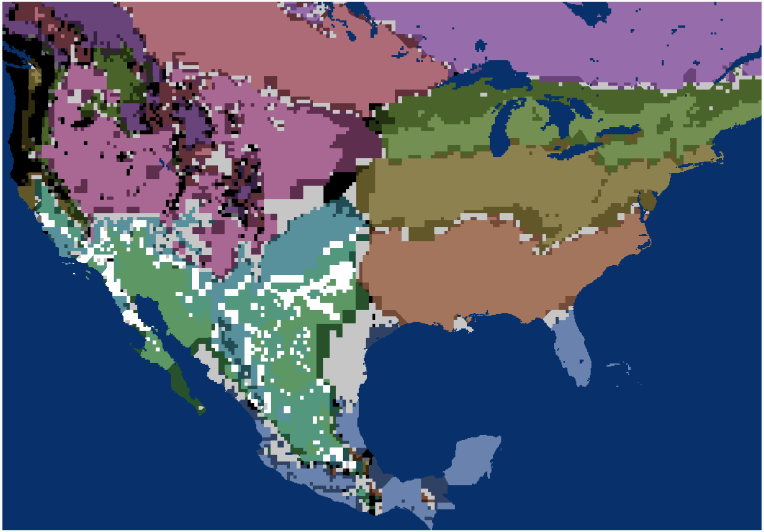
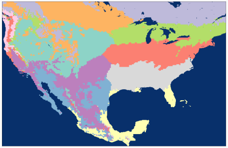


Figure 15: Consensus clustering from reduced ensemble of clusters for k=10, along with the trust. Grey = multi-class. Darker hue = lower trust.

Conclusion

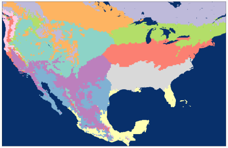
Summary

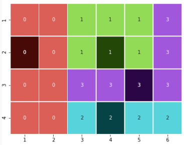
- The DWT brings forth structure hidden at different scales within the data.



Summary

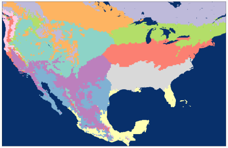
- The DWT brings forth structure hidden at different scales within the data.
- Mutual information allows us to effectively represent the diversity across all scales.

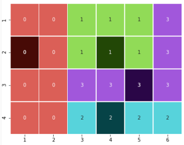


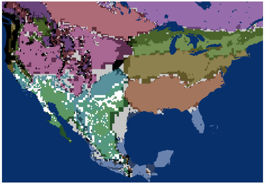


Summary

- The DWT brings forth structure hidden at different scales within the data.
- Mutual information allows us to effectively represent the diversity across all scales.
- Using this reduced ensemble, we produce a fuzzy clustering that has an interpretable trust metric at each point in space.







Extra

Extra

Mutual Information

- Let $U = \{U_j\}_{j=1}^k, V = \{V_j\}_{j=1}^l$ be two partitions of the data $X = \{x_i\}_{i=1}^n$.
- *Entropy* $\mathcal{H}(U)$ is average information (e.g., bits) needed to encode the cluster label for each data points of U.
- The *conditional entropy* $\mathcal{H}(U|V)$ denotes the average amount of information needed to encode U if V is known.
- **Mutual Information** $\mathcal{I}(U, V)$ measures how knowledge of one clustering reduces our uncertainty of the other:

$$\mathcal{I}(U, V) = \mathcal{H}(U) - \mathcal{H}(U|V).$$

- Assume points of X are sampled uniformly. Then,
  1. probability $x \in X$ in cluster U_i is $p(x) = \frac{|U_i|}{n}$
  2. probability $x, y \in X$ satisfy $x \in U_i$, $y \in V_j$ is $p(x, y) = \frac{|U_i \cap V_j|}{n}$
- We normalize mutual information:

$$\mathcal{NI}(U, V) := \frac{2\mathcal{I}(U, V)}{\mathcal{H}(U) + \mathcal{H}(V)}.$$

Extra

Results - Effect of k

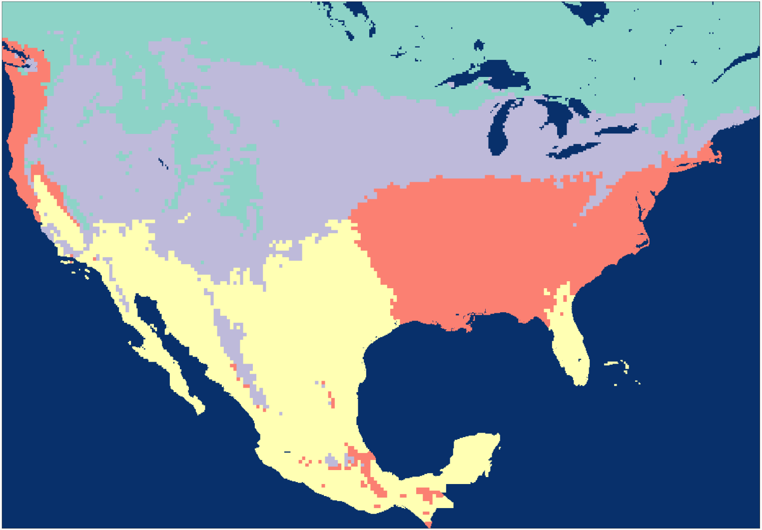


Figure 16: CGC: K-means $k = 4$, $(\ell_s, \ell_t) = (2, 3)$

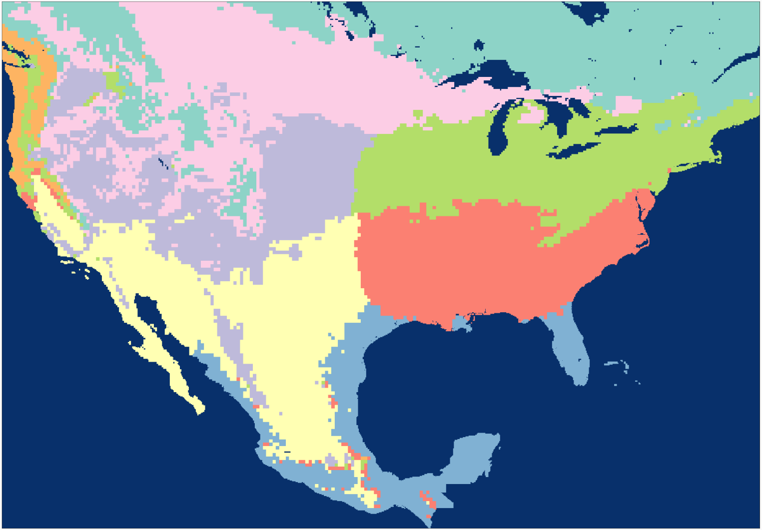


Figure 17: CGC: K-means $k = 8$, $(\ell_s, \ell_t) = (2, 3)$

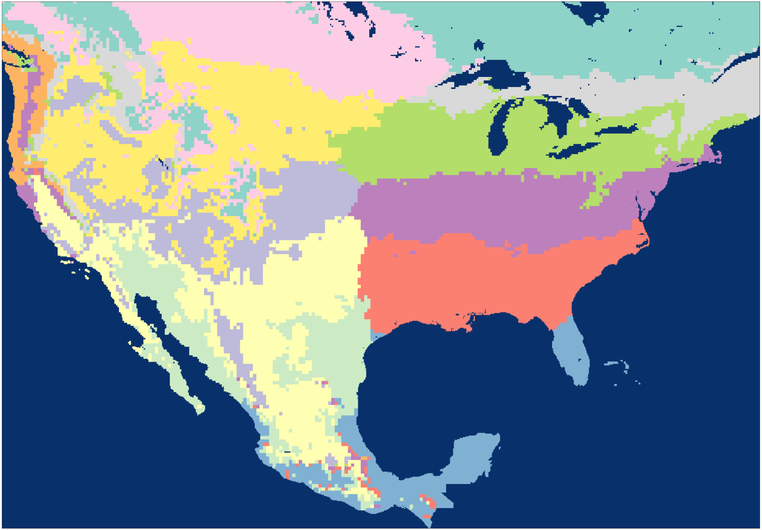


Figure 18: CGC: K-means $k = 12$, $(\ell_s, \ell_t) = (2, 3)$

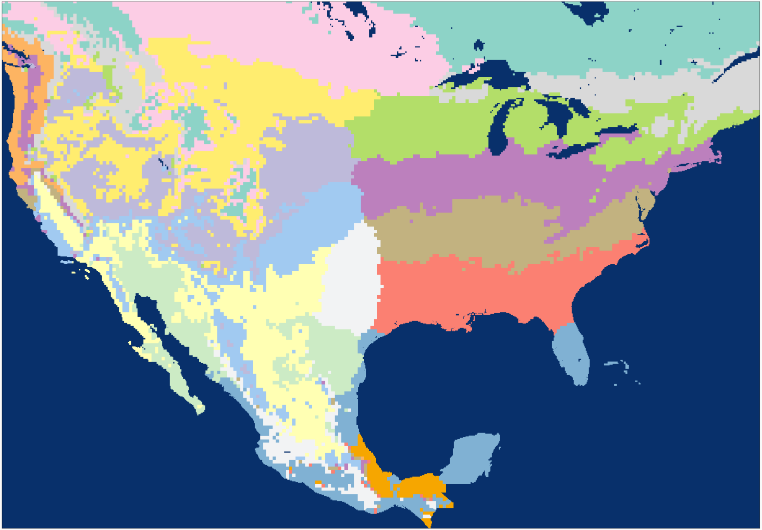


Figure 19: CGC: K-means $k = 16$, $(\ell_s, \ell_t) = (2, 3)$