ML4CC: Lecture 6

Sit with your discussion groups (same as last time)!

Assignments reminder

Keep doing your weekly PMIRO+Q

Your second coding assignment was due today before the start of class.

Your third coding assignment will be due Friday, March 15th by 11:59pm.

You have an exam on March 14th (8am)

Recap of previous paper

P: No way to say if 2 atmospheric states are “similar”

M: Use self-supervised learning (temporal difference prediction) to learn representation of atmospheric states, and use this representation as the basis of a distance metric (AtmoDist)

I: Self-supervised learning for this problem

R: AtmoDist behaves intuitively and better than image-based losses in tasks such as super-resolution

O: Does this definition of similarity capture what matters for climate models (e.g. physical laws)?

Climate Change in the News

CLIMATE

Wildfires are killing California's ancient giants. Can seedlings save the species?

FEBRUARY 26, 2024 · 5:00 AM ET

By Lauren Sommer, Ryan Kellman

Over only two years, about one-fifth of all giant sequoias have been killed in extreme wildfires in California. The numbers shocked ecologists, since the enormous trees can live more than 2,000 years and have evolved to live with frequent, low-intensity fires in the Sierra Nevada.

The smaller numbers of seedlings concerned scientists and the National Park Service. So in a historic step, the agency for the first time has begun replanting some severely burned areas. With a life span of thousands of years, the new seedlings will grow up in a climate that's rapidly changing. So, park officials are bringing in seedlings from other sequoia groves, ones that may have the genetic tools to handle a more hostile future.

The debate is one occurring on public lands across the country as the impacts of climate change get worse. Land managers face a key question: As humans take an increasing toll on natural landscapes, how far should we go to fix it?

Paper 5 Discussion

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2022.

Towards Global Crop Maps with Transfer Learning

Hyun-Woo Jo [*1] **Alkiviadis Koukos** [*2] **Vasileios Sitokonstantinou**[2]
Woo-Kyun Lee[1] **Charalampos Kontoes**[2]
[1] Department of Environmental Science and Ecological Engineering, Korea University
[2]BEYOND Centre, IAASARS, National Observatory of Athens
{akoukos,vsito,kontoes}@noa.gr
endeavor4a1@gmail.com
leewk@korea.ac.kr

Abstract

The continuous increase in global population and the impact of climate change on crop production are expected to affect the food sector significantly. In this context, there is need for timely, large-scale and precise mapping of crops for evidence-based decision making. A key enabler towards this direction are new satellite missions that freely offer big remote sensing data of high spatio-temporal resolution and global coverage. During the previous decade and because of this surge of big Earth observations, deep learning methods have dominated the remote sensing and crop mapping literature. Nevertheless, deep learning models require large amounts of annotated data that are scarce and hard-to-acquire. To address this problem, transfer learning methods can be used to exploit available annotations and enable crop mapping for other regions, crop types and years of inspection. In this work, we have developed and trained a deep learning model for paddy rice detection in South Korea using Sentinel-1 VH time-series. We then fine-tune the model for i) paddy rice detection in France and Spain and ii) barley detection in the Netherlands. Additionally, we propose a modification in the pre-trained weights in order to incorporate extra input features (Sentinel-1 VV). Our approach shows excellent performance when transferring in different areas for the same crop type and rather promising results when transferring in a different area and crop type.

Attendance

Select one person from the group to go to this Google Doc and write down the names of all people present in the group (remember to mark who took attendance!). **If someone is virtual, mark it with a V**.

https://docs.google.com/document/d/1PKhw9E2lJpAnFrFO88DOc2rscZFVlcIv47QNa5h1sGs/edit?usp=sharing (link is in Brightspace under Syllabus content)

Discussion Question 1

What are the dimensions of the input and output of this network?

Sequence of images to single crop mask

Input is a length-8 sequence of 256x256x1 SAR images.

The output is a 256x256x1 segmentation map representing rice/not rice.

A.1 Model Details

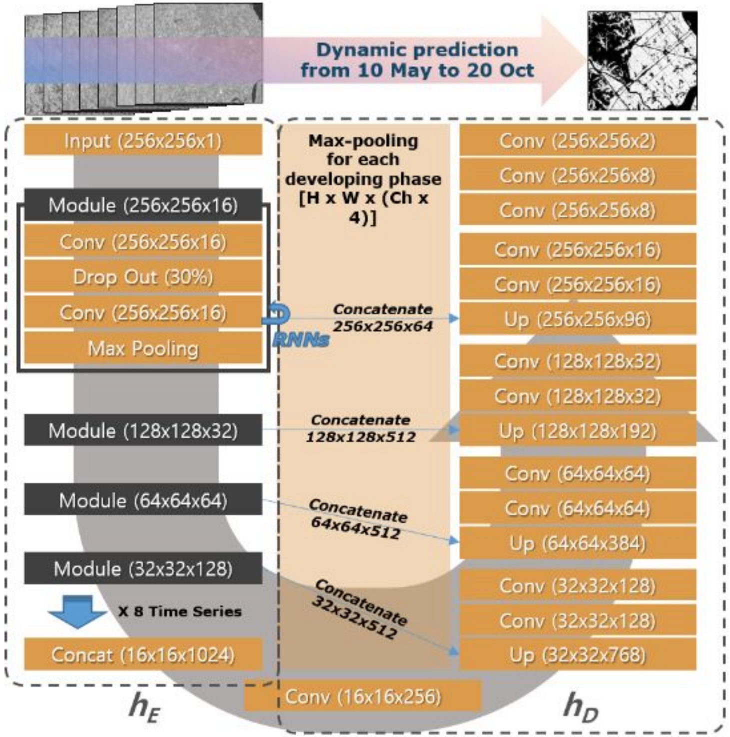


Figure 2: Recurrent U-net architecture

Discussion Question 2

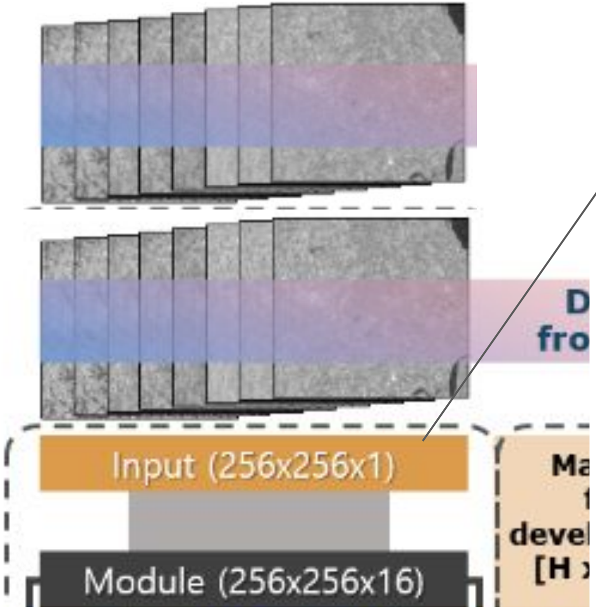
What are the authors doing here and why?

Incorporation of additional feature types. In crop classification, diverse characteristics of each crop (e.g., texture, reflection) raise the need of an extended application of TL, such as using different sources of data as input. In this direction, we adapt h^p, pre-trained on Sentinel-1 VH backscatter, to take as input both Sentinel-1 VH and VV features. To do this, the pre-trained weights at the first layer of the encoder $(W^P_{E_0})$ are divided by the total number of input layers (Eq. $\boxed{3}$). Therefore, a similar scale of signal intensity is transferred to the activation functions (σ) that is invariant to the number of inputs, and ensures that h^p maintains the trained feature extraction process.

$$h^P_{E_0} = \sigma((W^P_{E_0} \cdot x^s + W^P_{E_0} \cdot x^{s'})/2 + b) \tag{3}$$

Adding inputs to pre-trained models

In general, it is easier to modify the output of a pre-trained model than the input. But here, they try to add another input to the model (the VV backscatter).



The weights learned here are only for one input channel!

Hack: To have weights for the new channel, they just copy the weights for the initial channel, then divide all the weights by two. Now the model gets the same magnitude of overall input, but from two channels instead of one. (input is now 256x256x2)

Discussion Question 3

For each of the marked scores, explain what region the corresponding model was pre-trained on, what region it was fine-tuned on, what region the performance is being reported for, and what crop type it is detecting.

Table 1: Mean IoU for the different scenarios and methods

Fine-tuning	Spain				France				The Netherlands	
Test	Spain		France		France		Spain		The Netherlands	
Feature	VH	VH|VV	VH	VH|VV	VH	VH|VV	VH	VH|VV	VH	VH|VV
RF	0.87	**0.90**	**0.63**	**0.66**	0.76	0.84	0.77	0.78	0.26	0.40
RI	0.86	0.69	0.52	0.36	0.76	0.74	0.70	0.73	0.31	0
FT	0.89	**0.90**	0.57	0.63	0.82	0.83	**0.82**	0.83	0.40	0.45
FT_E	**0.90**	**0.90**	**0.63**	**0.66**	**0.86**	**0.86**	0.79	**0.84**	**0.42**	**0.54**

For each of the marked scores, explain what region the corresponding model was pre-trained on, what region it was fine-tuned on, what region the performance is being reported for, and what crop type it is detecting.

Korea, Spain, France, Rice

Not pre-trained, France, France, Rice

Korea, Netherlands, Netherlands, Barley

Table 1: Mean IoU for the different scenarios and methods

Fine-tuning	Spain				France				The Netherlands	
Test	Spain		France		France		Spain		The Netherlands	
Feature	VH	VH|VV	VH	VH|VV	VH	VH|VV	VH	VH|VV	VH	VH|VV
RF	0.87	**0.90**	**0.63**	**0.66**	0.76	0.84	0.77	0.78	0.26	0.40
RI	0.86	0.69	0.52	0.36	0.76	0.74	0.70	0.73	0.31	0
FT	0.89	**0.90**	0.57	0.63	0.82	0.83	**0.82**	0.83	0.40	0.45
FT_E	**0.90**	**0.90**	**0.63**	**0.66**	**0.86**	**0.86**	0.79	**0.84**	**0.42**	**0.54**

Discussion Question 4

The authors compare their pre-trained Recurrent U-net model to Random Forest models that were trained on data from France and Spain. Why were they able to train the Random Forest models, but not the U-net model, on this data? For which region(s) does using the pre-trained U-net help the performance?

Comparison to “local” models

Random Forest models tend to have fewer parameters and therefore require less data to train than deep neural networks.

RF models trained on data from France and the Netherlands perform worse than the pre-trained U-net

Table 1: Mean IoU for the different scenarios and methods

Fine-tuning	Spain				France				The Netherlands	
Test	Spain		France		France		Spain		The Netherlands	
Feature	VH	VH|VV	VH	VH|VV	VH	VH|VV	VH	VH|VV	VH	VH|VV
RF	0.87	**0.90**	**0.63**	**0.66**	0.76	0.84	0.77	0.78	0.26	0.40
RI	0.86	0.69	0.52	0.36	0.76	0.74	0.70	0.73	0.31	0
FT	0.89	**0.90**	0.57	0.63	0.82	0.83	**0.82**	0.83	0.40	0.45
FT_E	**0.90**	**0.90**	**0.63**	**0.66**	**0.86**	**0.86**	0.79	**0.84**	**0.42**	**0.54**

Discussion Question 5

What does this figure tell us about differences between rice in Spain and France?

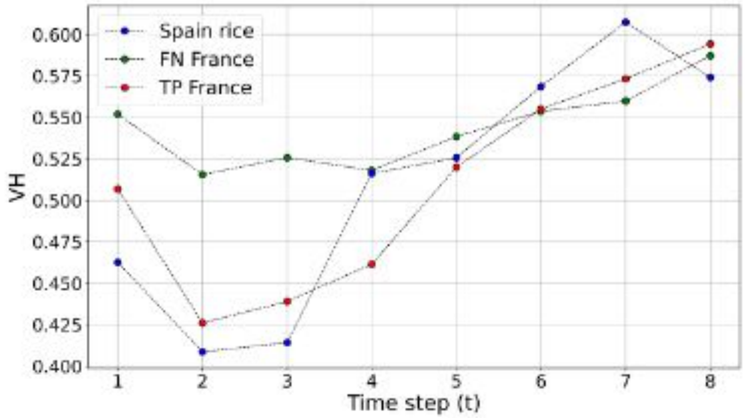


Figure 1: Green and red dots represent the mean VH time-series of the True Positive (TP) and False Negative (FN) predictions of the reccurent U-net fine-tuned in Spain and tested in France. The blue dots represent the mean VH time-series of rice instances in Spain

Two types of French rice

France seems to have two different types of rice, one that is similar to the rice in Spain (and produces a large increase in VH) and one that is not (and has somewhat steady VH). The model fine-tuned on Spain doesn't capture the second kind (but the French fine-tuned model can presumably capture it, and does a decent job of identifying rice in Spain).

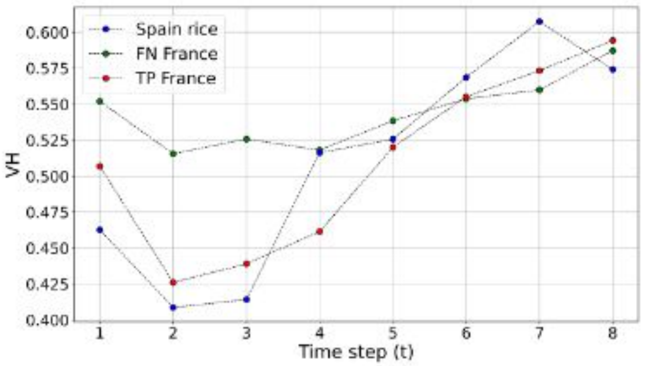


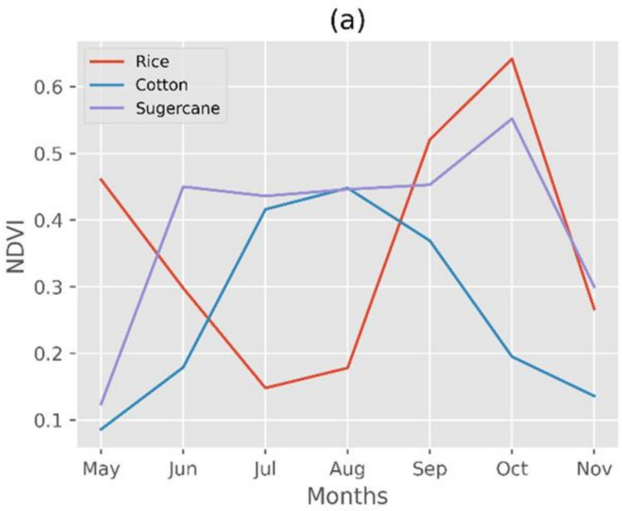
Figure 1: Green and red dots represent the mean VH time-series of the True Positive (TP) and False Negative (FN) predictions of the reccurent U-net fine-tuned in Spain and tested in France. The blue dots represent the mean VH time-series of rice instances in Spain

Discussion Question 6

Is rice detection more or less difficult than barley detection? Why?

Flooding can make rice detection easy

Transferring the paddy rice model to predict summer barley does not perform as well. Paddy rice fields are intentionally flooded at the start of the cultivation period; SAR data have a great ability of identifying water content, which makes them ideal in classifying paddy rice. However, this is not the case for summer barley, and thus the discrimina-



Flooding or transplanting phase		High vegetation period			Harvesting period	
May	June	July	August	September	October	November

Table 1. Rice crop calendar in southern Punjab, Pakistan.

Waleed et al., 2022

Other crops don’t necessarily have quite as strong of a signal

Discussion Question 7

What is the difference between the FT_E , $FT_{D,}$ and FT models? Which performed best and which performed worst?

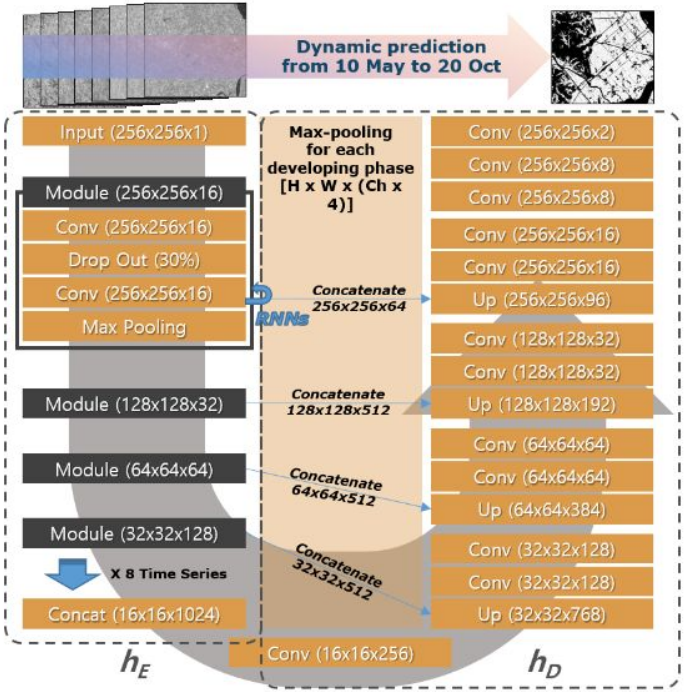
Which weights to fine-tune

FT_E lets the encoder weights be fine-tuned but freezes the decoder weights

FT_D lets the decoder weights be fine-tuned but freezes the encoder weights

FT lets all weights be fine tuned

A.1 Model Details



Based on our experiments, we found FT_E achieved better performance than FT and FT_D. Fine-tuning the decoder did not converge, whereas by fine-tuning only its last (or 2-3 last) layers the model was successfully trained, but provided suboptimal performance. Table 1 presents the Intersection over

Discussion Question 8

Share what questions you wrote in your PMIRO+Q and decide as a group what you'd like to ask.

Update your PMIRO+Q

Submit a second file to the Brightspace assignment (don’t overwrite the original):

It should:

Update your PMIRO as needed

Answer your own Q

You can be talking with your group during this!

15 min break

Lecture

Climate Change: Human beliefs and how to change them

Machine Learning: Natural Language Processing, Transformer, Topic Clustering

Everything is done by people

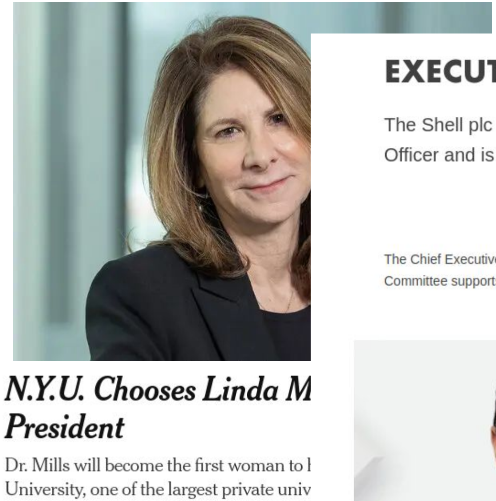
N.Y.U. Chooses Linda Mills as Its Next President

Dr. Mills will become the first woman to head New York University, one of the largest private universities in the country.

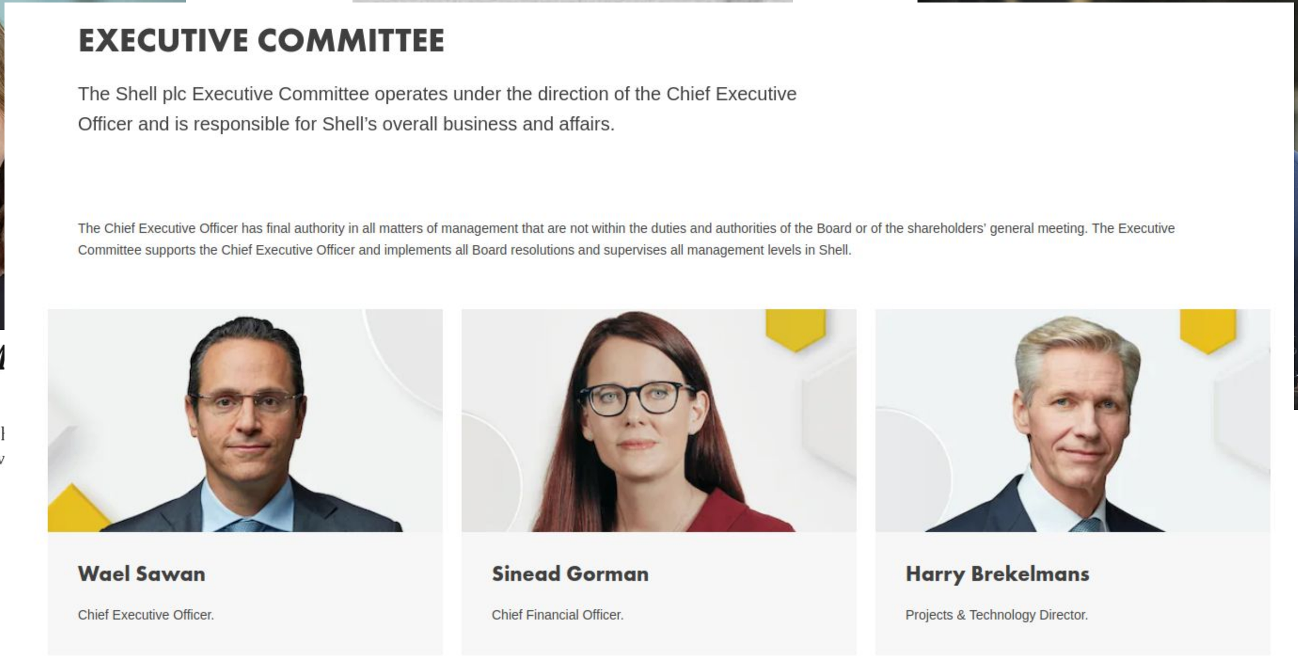
Stephen John Brademas Jr. (March 2, 1927 – July 11, 2016). NYU President from 1981 to 1992

Eric Adams, Mayor of New York City since 2022

Everything is done by people


N.Y.U. Chooses Linda M
President

Dr. Mills will become the first woman to l
University, one of the largest private univ


EXECUTIVE COMMITTEE

The Shell plc Executive Committee operates under the direction of the Chief Executive Officer and is responsible for Shell's overall business and affairs.

The Chief Executive Officer has final authority in all matters of management that are not within the duties and authorities of the Board or of the shareholders' general meeting. The Executive Committee supports the Chief Executive Officer and implements all Board resolutions and supervises all management levels in Shell.

Wael Sawan

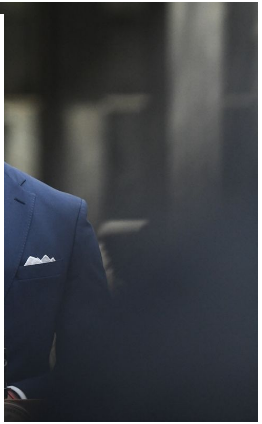
Chief Executive Officer.

Sinead Gorman

Chief Financial Officer.

Harry Brekelmans

Projects & Technology Director.

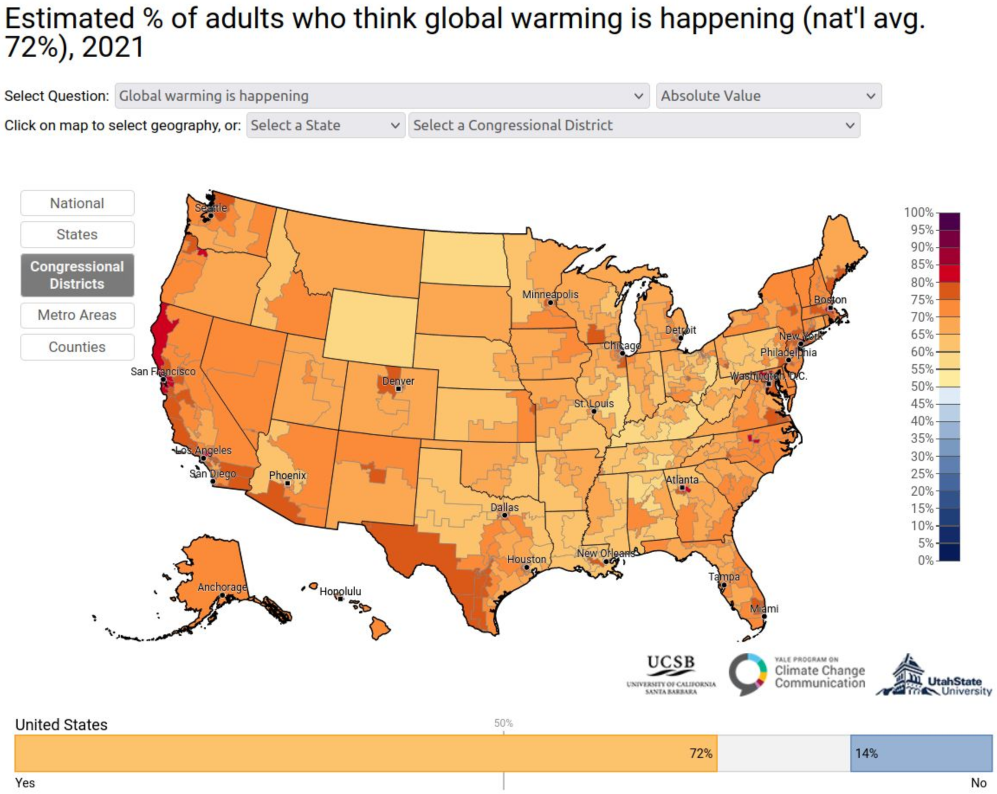


New York

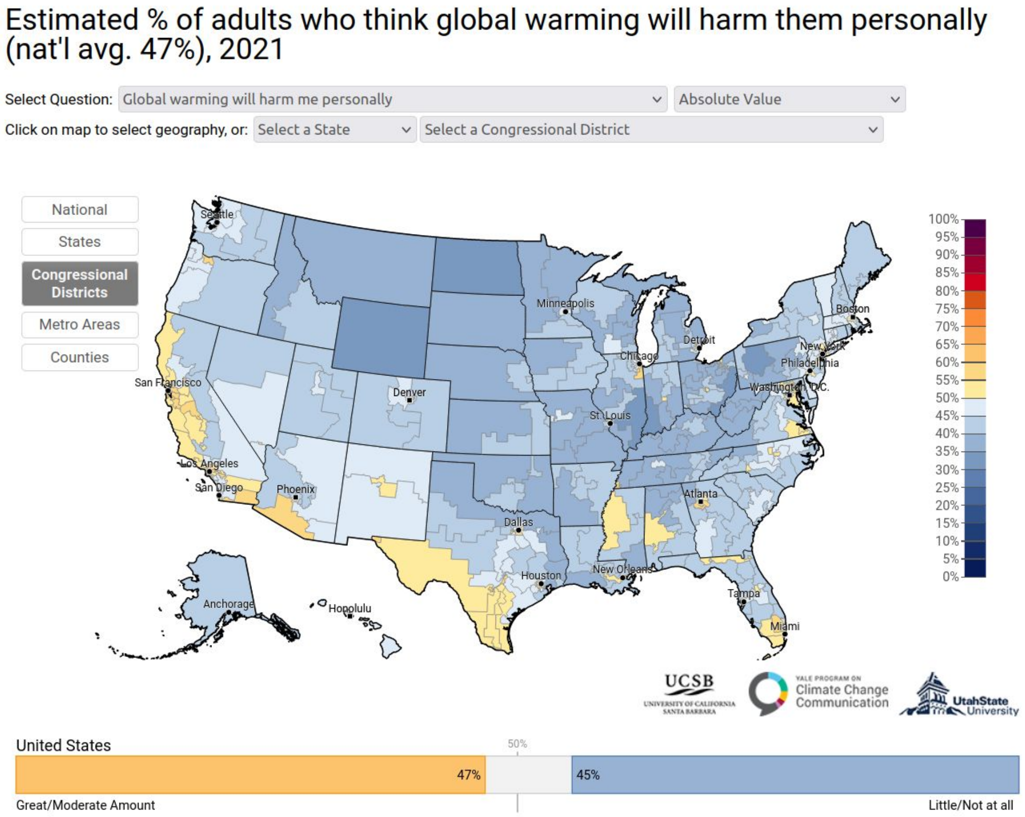
Everything is done by people

Not just as individuals, but also collectively

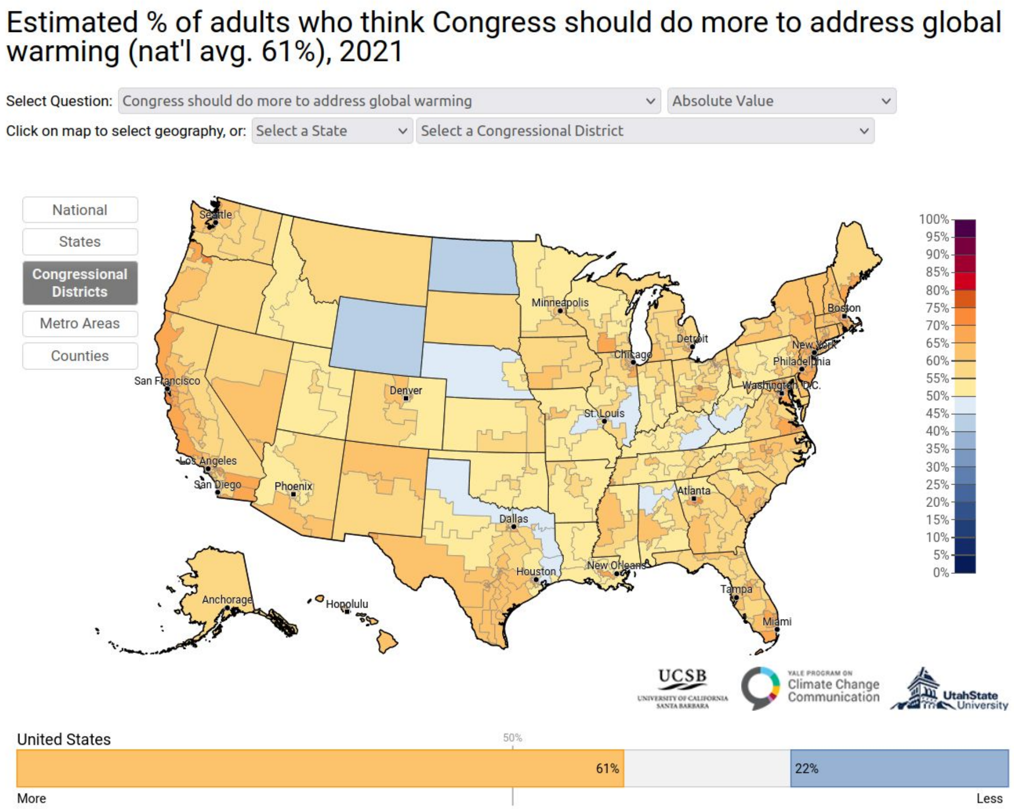
What do people think about climate change?



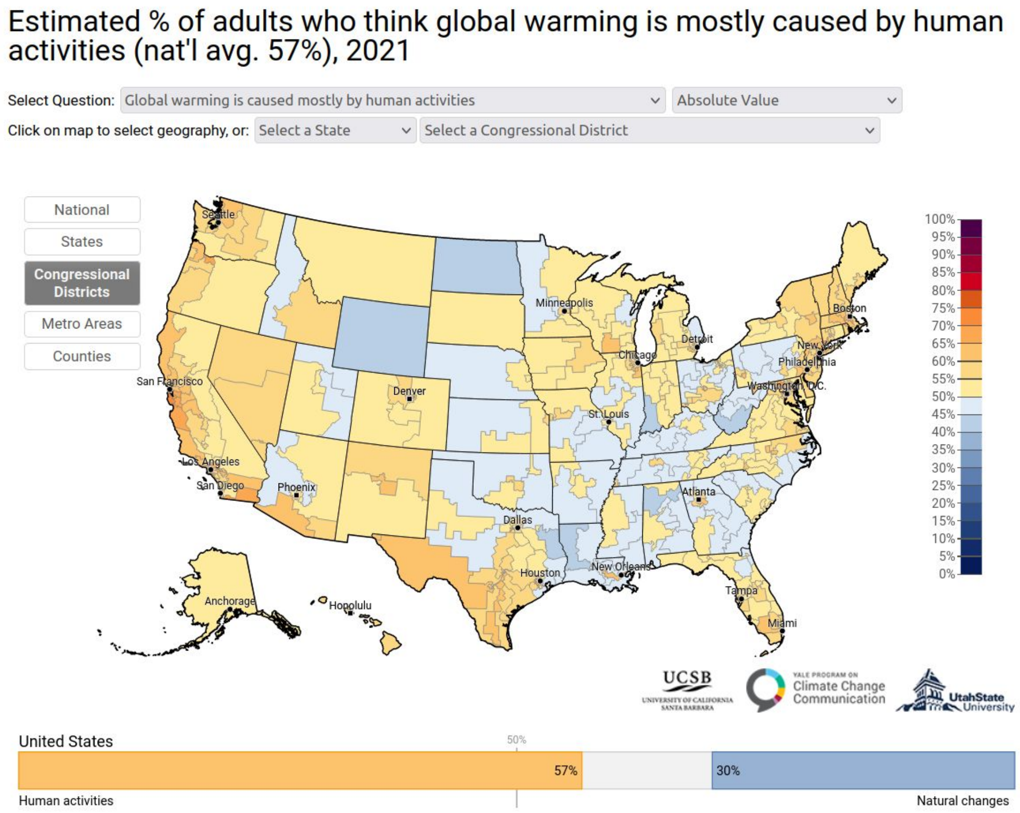
What do people think about climate change?



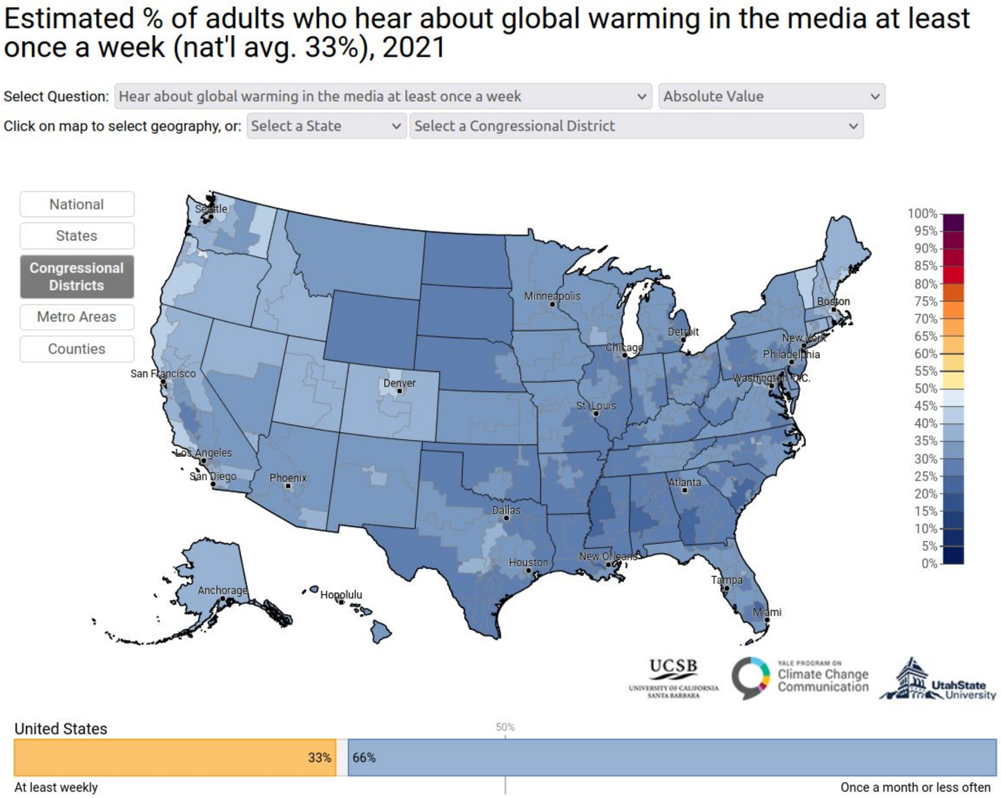
What do people think about climate change?



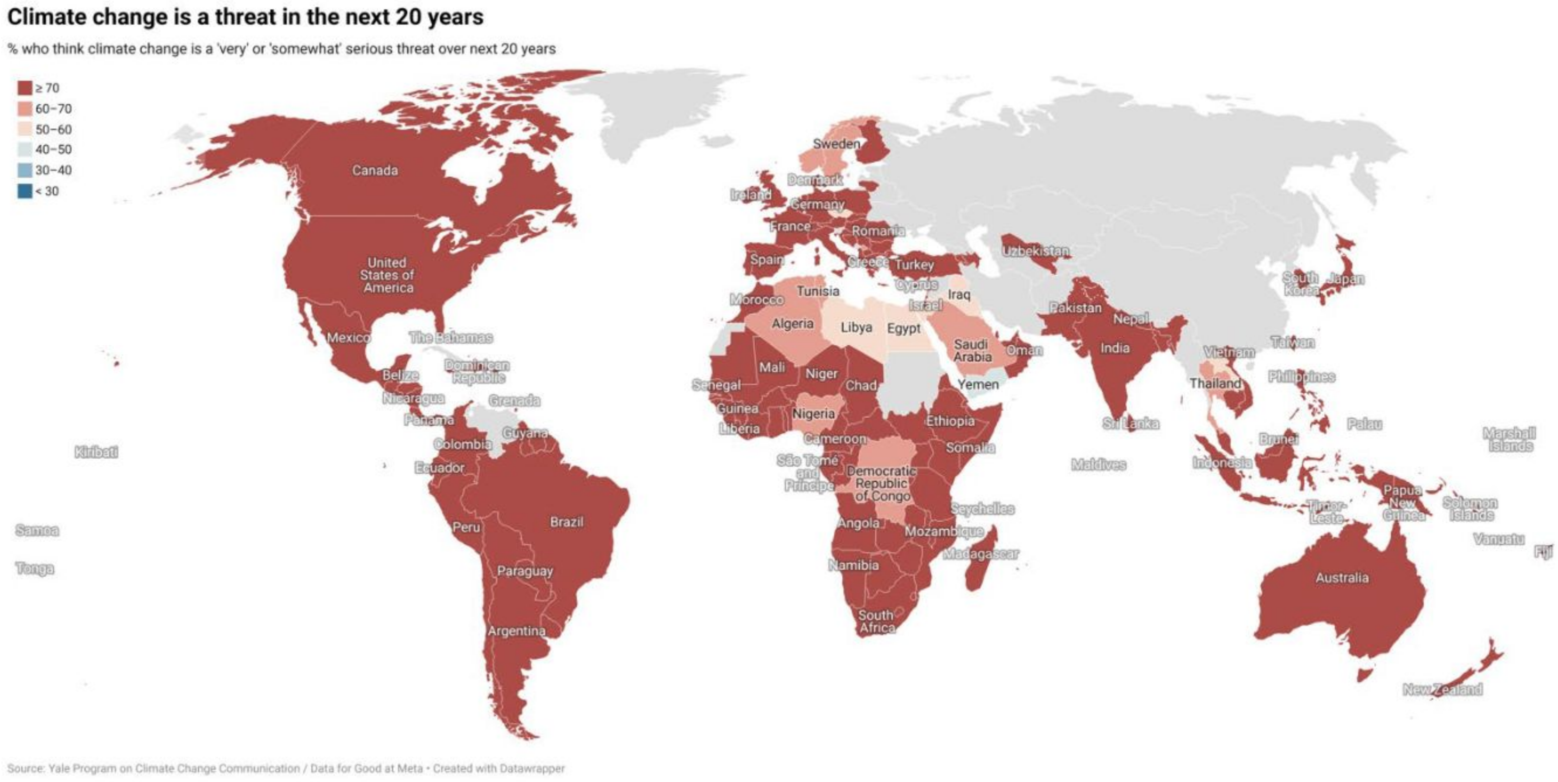
What do people think about climate change?



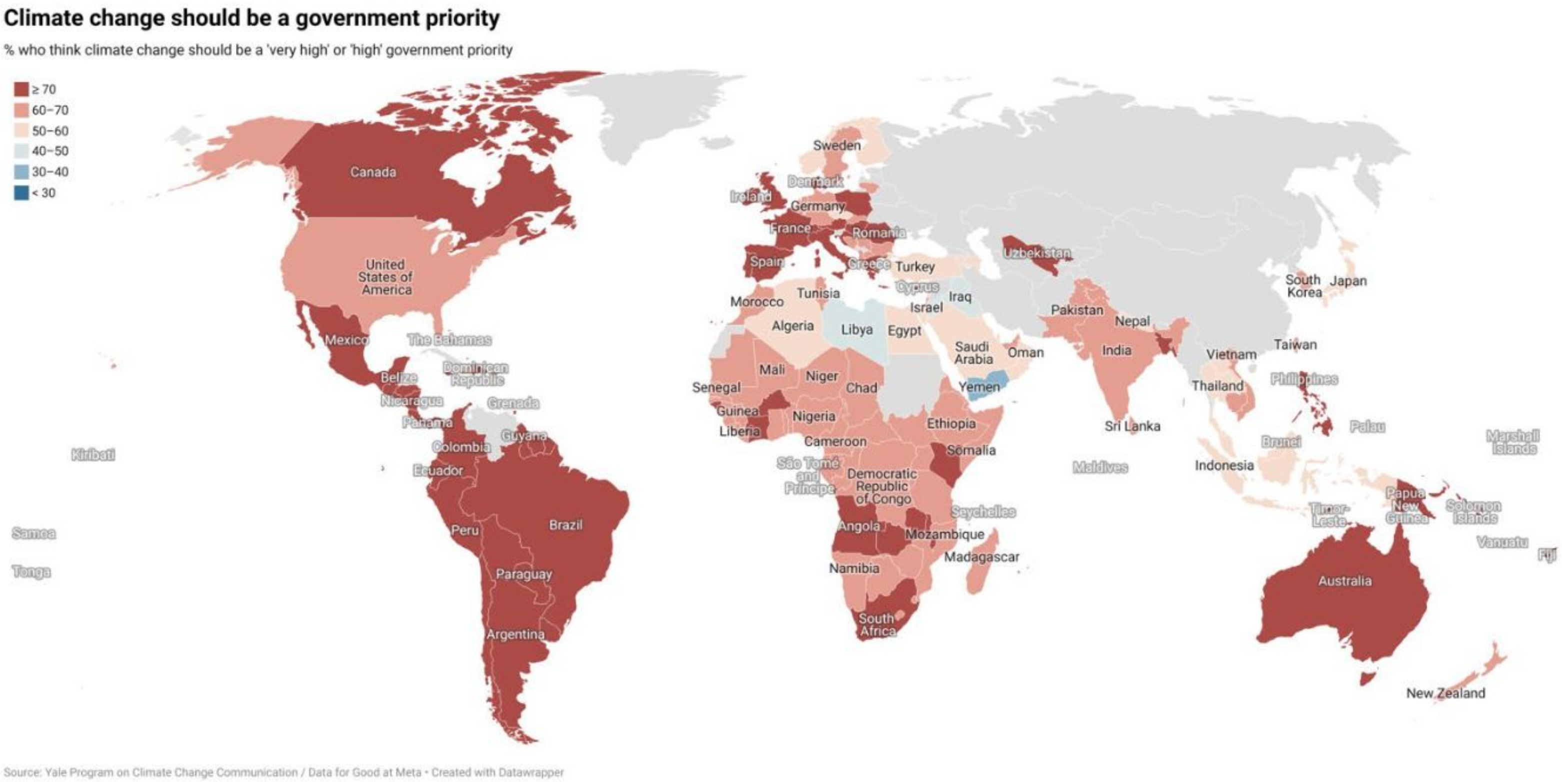
What do people think about climate change?



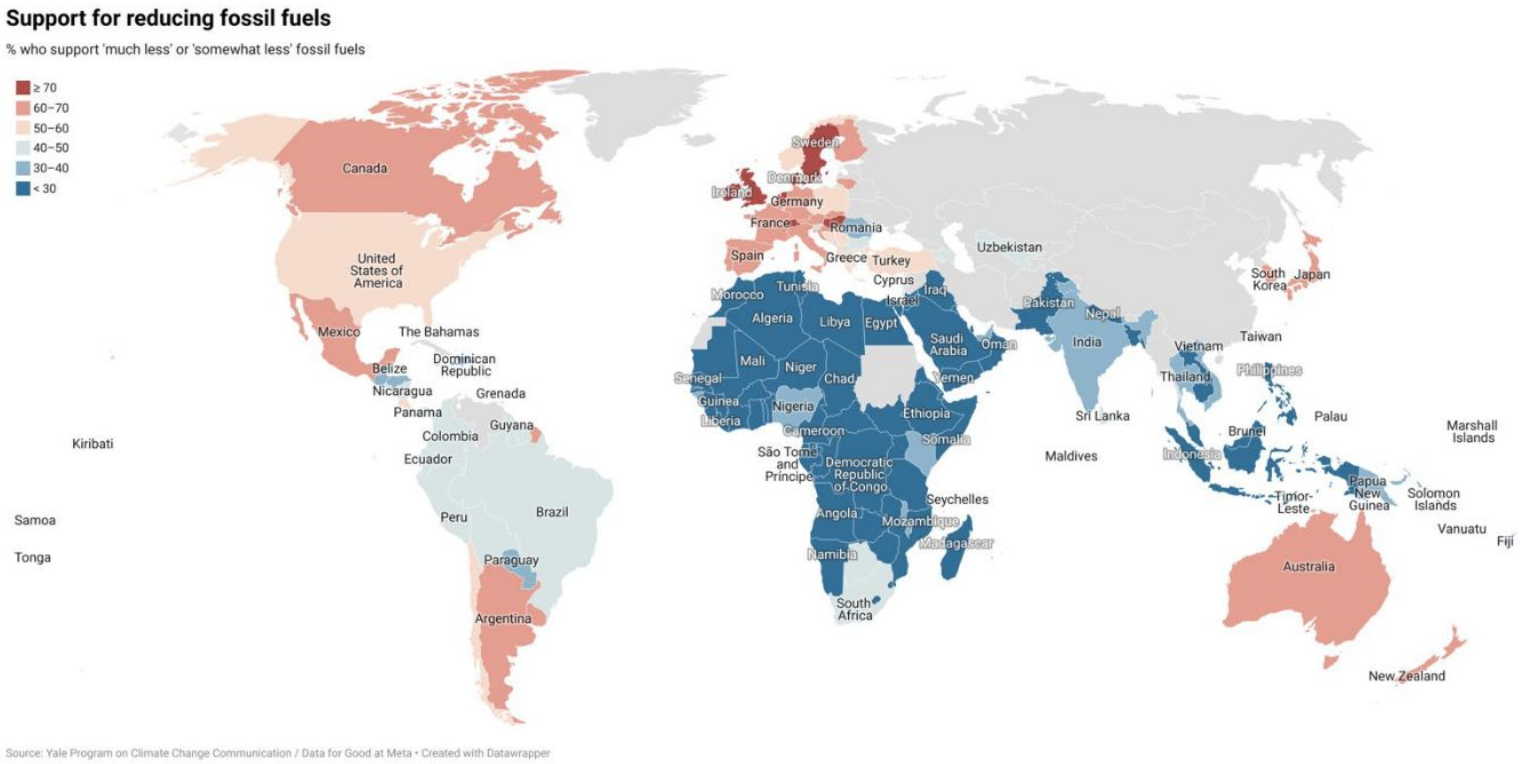
What do people think about climate change?



What do people think about climate change?



What do people think about climate change?



Can we change people's minds?

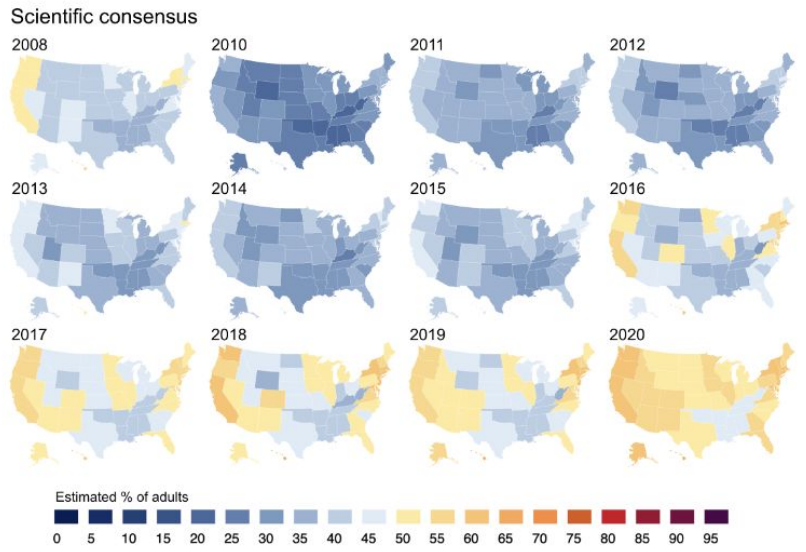


Figure 1. Variations in percentage of Americans by state and over time who think that most scientists think that global warming is happening, 2008–2020.

It's already been happening

https://iopscience.iop.org/article/10.1088/1748-9326/aca702/meta

Can we change people’s minds?

Oil companies have done it (in the wrong way)

Who told you the earth was warming... Chicken Little?

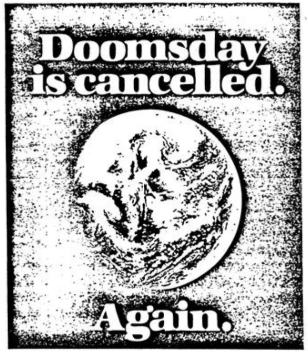
Chicken Little's hysteria about the sky falling was based on a fact that got blown out of proportion.

It's the same with global warming. There's no hard evidence it is occurring. In fact, evidence the Earth is warming is weak. Proof that carbon dioxide has been the primary cause is non-existent. Climate models cannot accurately predict far-future global change. And the underlying physics of climatic change are still wide open to debate.

If you care about the earth, but don't want your imagination to run away with you, make sure you get the facts.

Write Informed Citizens for the Environment, P.O. Box 1513, Grand Forks, North Dakota 58206, or call toll-free 1-701-746-4573. We'll send today's

Informed Citizens



The twentieth century has seen many predictions of global destruction. In the 1930's, some scientists claimed we were in the middle of a disastrous warming trend. In the mid 1970's, others were sure we were entering a new Ice Age. And so on.

It's the same with global warming. There's no hard evidence it is occurring. In fact, evidence the Earth is warming is weak. Proof that carbon dioxide has been the primary cause is non-existent. Climate models cannot accurately predict far-future global change. And the underlying physics of the climatic change are still wide open to debate.

If you care about the environment, but don't care to be pressured into spending money on problems that don't exist, make sure you get the facts.

Write: Informed Citizens for the Environment, P.O. Box 1513, Grand Forks, North Dakota 58206 or call (701) 746-4573. We'll send you the facts about global warming.

Informed Citizens for the Environment

Lies they tell our children

"I don't have a future."

With tears streaming down her face, a 13-year-old girl made this bleak assessment to her father. To back up her pessimism, she had brought home from school a mimeographed sheet listing the horrors that awaited her generation in the next 25 years: Worldwide famine, overpopulation, air pollution so bad that everyone would wear a gas mask, befouled rivers and streams that would mandate cleansing tablets in drinking water... a greenhouse effect that would melt the polar ice caps and devastate U.S. coastal cities... a cancer epidemic brought on by damage to the ozone layer.

Moved by the girl's misery, her father, Herbert I. London of the Hudson Institute and New York University, wrote a book, *Why Are They Lying to Our Children?* The book documents how some of the myths of the 1960s and 1970s—and some much older than that—are being perpetuated and taught as gospel truth in some of our schools. And the book raises a question in our minds: Will the next generation have any better understanding of science and technology—both their merits and their problems—than our own?

Professor London's book is not a plea for unbridled technology. But it is a plea for balance. And school textbooks, he believes, are notoriously unbalanced. In dealing with environmental questions, for example, no textbook the professor could find made any mention of the following facts:

- Total automobile emissions of hydrocarbons, carbon monoxide, and nitrogen oxide in the U.S. are less than half what they were from 1957 to 1967.
- The amount of unhealthy sulfur dioxide in the air has been steadily declining since 1970.
- The bacteria level in the Hudson River declined by more than 30 percent between 1966 and 1980.

Textbooks, Professor London finds, mythologize nature as eternally benign until disturbed by man. It's a rare schoolbook that talks about volcanoes belching radiation into the air, floods that overwhelm river towns, and tornadoes that lift people into oblivion. Moreover, textbooks hardly mention the promise of a bright future already on the horizon—when average life expectancy may approach 90 years, when products derived from recombinant DNA research will eliminate most viral diseases, when we will enjoy greater leisure, and materials—especially plastics—will be better, stronger, and safer.

Professor London's conclusion—with which we heartily agree—is that we should help our children think for themselves and reach balanced conclusions. Let's look at their textbooks, not to censor them but to raise questions. Let's give them different points of view and help discuss them. That way we can educate a new generation of citizens who aren't scared by science, and who won't be swayed by old mythologies.

Our youngsters do have a future. We, and the schools, should help them look forward to it with hope, even as they prepare to deal with its problems.

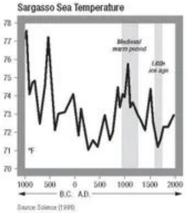
Mobil

Unsettled Science

Knowing that weather forecasts are reliable for a few days at best, we should recognize the enormous challenge facing scientists seeking to predict climate change and its impact over the next century. In spite of everyone's desire for clear answers, it is not surprising that fundamental gaps in knowledge leave scientists unable to make reliable predictions about future changes.

A recent report from the National Research Council (NRC) raises important issues, including these still-unanswered questions: (1) Has human activity already begun to change temperature and the climate, and (2) How significant will future change be?

The NRC report confirms that Earth's surface temperature has risen by about 1 degree Fahrenheit over the past 150 years. Some use this result to claim that humans are causing global warming, and they point to storms or floods to say that dangerous impacts are already under way. Yet scientists remain unable to confirm either contention.



Geological evidence indicates that climate and greenhouse gas levels experience significant natural variability for reasons having nothing to do with human activity. Historical records and current scientific evidence show that Europe and North America experienced a *medieval warm period* one thousand years ago, followed centuries later by a *little ice age*. The geological record shows even larger changes throughout Earth's history. Against this backdrop of large, poorly understood natural variability, it is impossible for scientists to attribute the recent small surface temperature increase to human causes.

Moreover, computer models relied upon by climate scientists predict that lower atmospheric temperatures will rise as fast as or faster than temperatures at the surface. However, only within the last 20 years have reliable global measurements of temperatures in the lower atmosphere been available through the use of satellite technology. These measurements show little if any warming.

Even less is known about the potential positive or negative impacts of climate change. In fact, many academic studies and field experiments have demonstrated that increased levels of carbon dioxide can promote crop and forest growth.

So, while some argue that the science debate is settled and governments should focus only on near-term policies—that is empty rhetoric. Inevitably, future scientific research will help us understand how human actions and natural climate change may affect the world and will help determine what actions may be desirable to address the long-term.

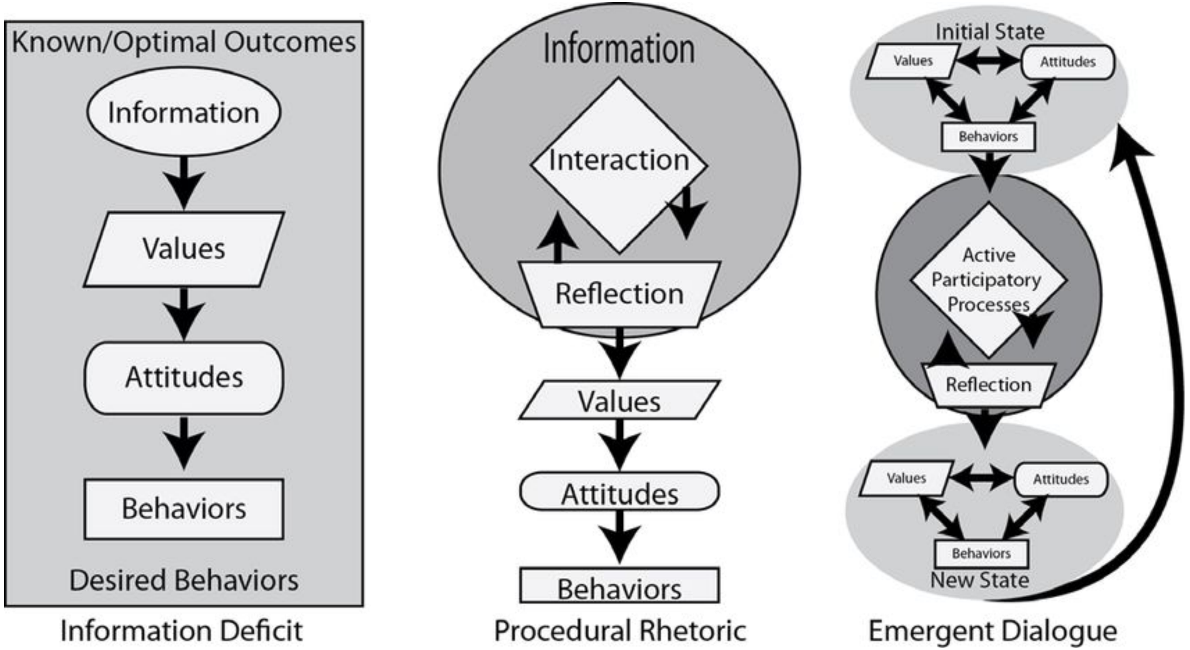
Science has given us enough information to know that climate changes may pose long-term risks. Natural variability and human activity may lead to climate change that could be significant and perhaps both positive and negative. Consequently, people, companies and governments should take responsible actions now to address the issue.

One essential step is to encourage development of lower-emission technologies to meet our future needs for energy. We'll next look at the promise of technology and what is being done today.

ExxonMobil

‘ExxonMobil’s climate “advertorials” – advertisements disguised as editorials – appeared in the op-ed page of the New York Times and other newspapers and were part of what scholars have called “the longest, regular (weekly) use of media to influence public and elite opinion in contemporary America”.’

What is effective at changing minds?



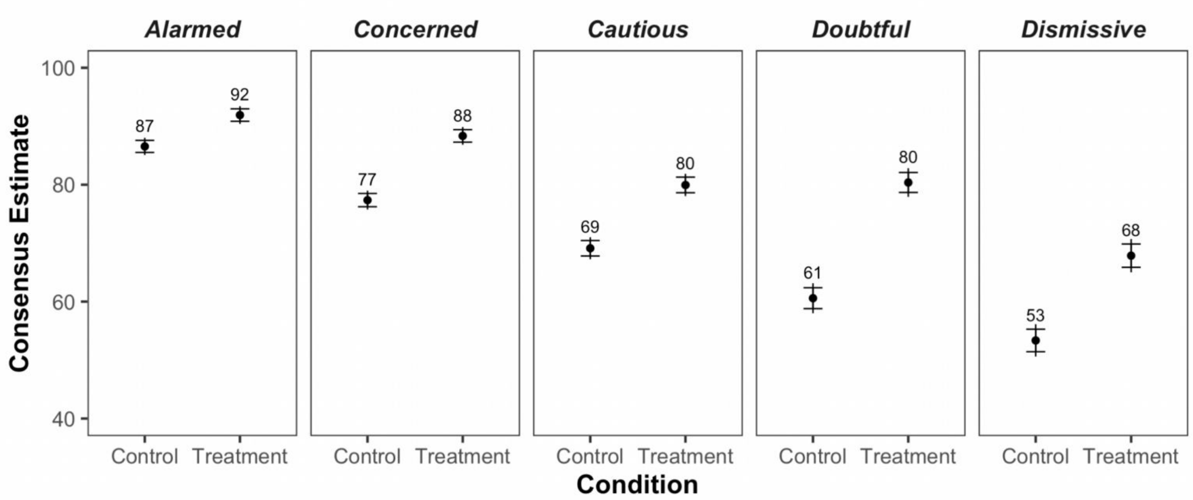
Ways to change minds depends on how you think minds work: The ability to change minds and behaviors can depend simply on providing information, or may require more involved processes of interaction and activations of a sense of identity

Tanenbaum et al., 2013

What is effective at changing minds?

Enforcing that there is scientific consensus

Estimates of the scientific consensus across conditions and audience segments



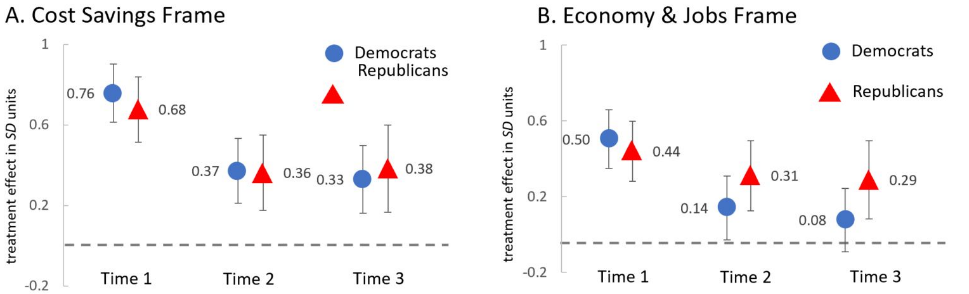
Note. Vertical error bars represent 95% confidence intervals. Horizontal error bars represent 83% confidence intervals to facilitate visual comparisons of significant differences at $p = .05$. Values are means adjusted for pre-treatment estimates of the scientific consensus.

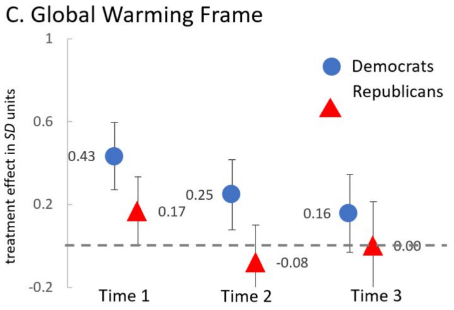
"We delivered a consensus message (i.e., "97% of climate scientists have concluded that human-caused global warming is happening") to members of five of the six U.S. climate audiences. We found that all audiences – from *Alarmed* to *Dismissive* – updated their beliefs about the scientific consensus."

https://climatecommunication.yale.edu/publications/communicating-the-scientific-consensus-on-climate-change-diverse-audiences-and-effects-over-time/

What is effective at changing minds?

Emphasizing co-benefits





The three panels show the effect of each of the three frames (Panel A = Cost Savings Frame; Panel B = Economy & Jobs Frame; Panel C = Global Warming Frame). The values in each panel represent the size of the effect (y-axis) of that frame on beliefs about that benefit of renewable energy, for Democrats and Republicans separately.

The x-axis shows how the size of these persuasive effects decayed over time. Time 1 measurement was immediately after viewing the message. Time 2 was an average of 11 days after Time 1. Time 3 was an average of 23 days after Time 1. Error bars indicate 95% confidence intervals around the mean.

University of CINCINNATI

YALE PROGRAM ON Climate Change Communication

What is effective at changing minds?

Changing actions can change beliefs

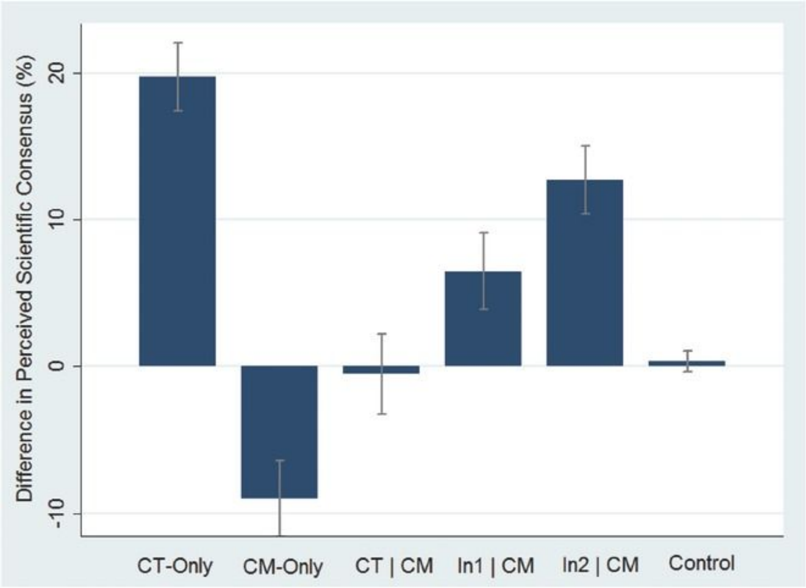
In fact, taking action with concrete solutions can actually help change minds. "Belief and action are connected," said anthropologist Ben Orlove, co-director of the Earth Institute's Center for Research on Environmental Decisions. "Belief is often a basis for action. But once you're committed to a course of action, you tend to find lots of reasons for why you did it."

Hayhoe told a story that illustrates just this point. For years, her colleague argued the science of climate change with his father who was a long-time doubter, but he was never able to change his father's mind. Finally the local community offered a big rebate to get solar panels, so the father installed them on his house. One year later, after telling everyone what a good deal it was and how much money he had saved, the father came to Hayhoe's colleague and said, "You know, that climate thing might be real after all."

https://news.climate.columbia.edu/2017/08/09/what-changes-minds-about-climate-change/

What is effective at changing minds?

‘Inoculation’ against known misinformation



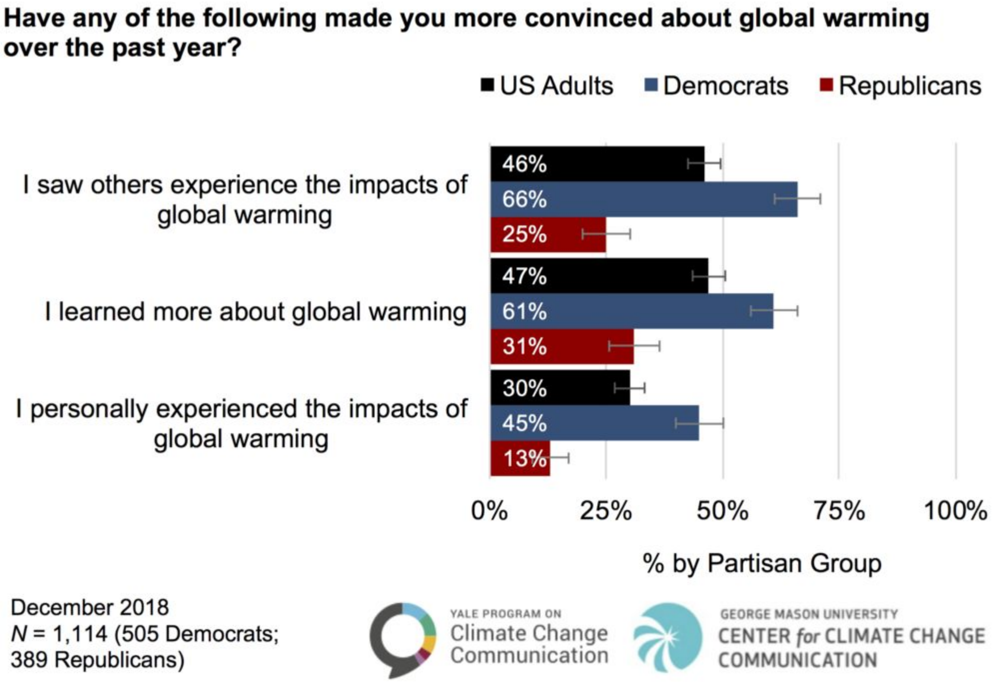
Note: CT = Consensus Treatment, CM = Counter-Message, In1 = General Inoculation, In2 = Detailed Inoculation. Error bars represent 95% confidence intervals.

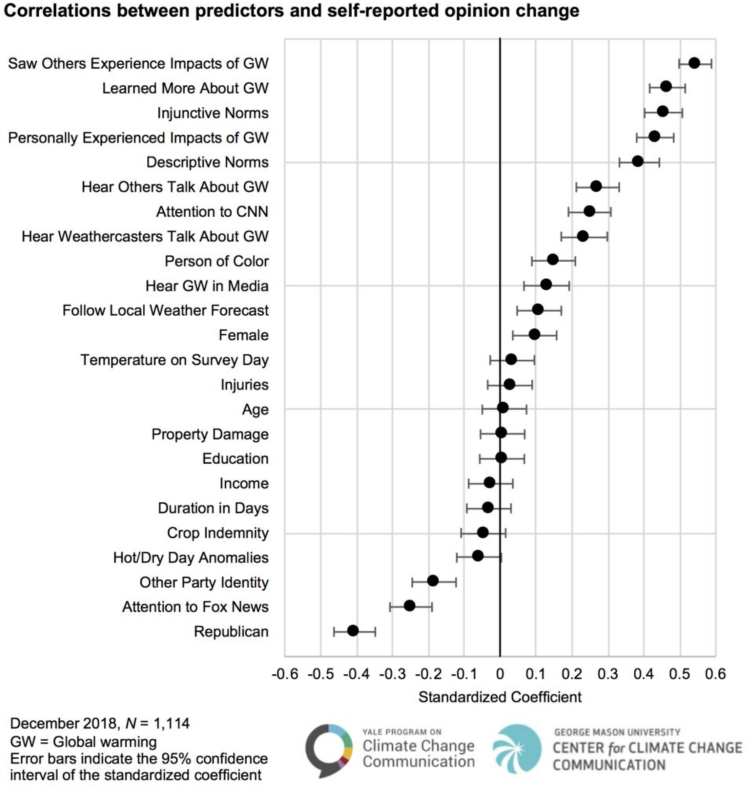
Letting people know that politically-motivated actors are spreading misinformation about climate change (In1 and In2) can reduce the impact of that misinformation.

https://onlinelibrary.wiley.com/doi/full/10.1002/gch2.201600008

What is effective at changing minds?

Experience with the effects of climate





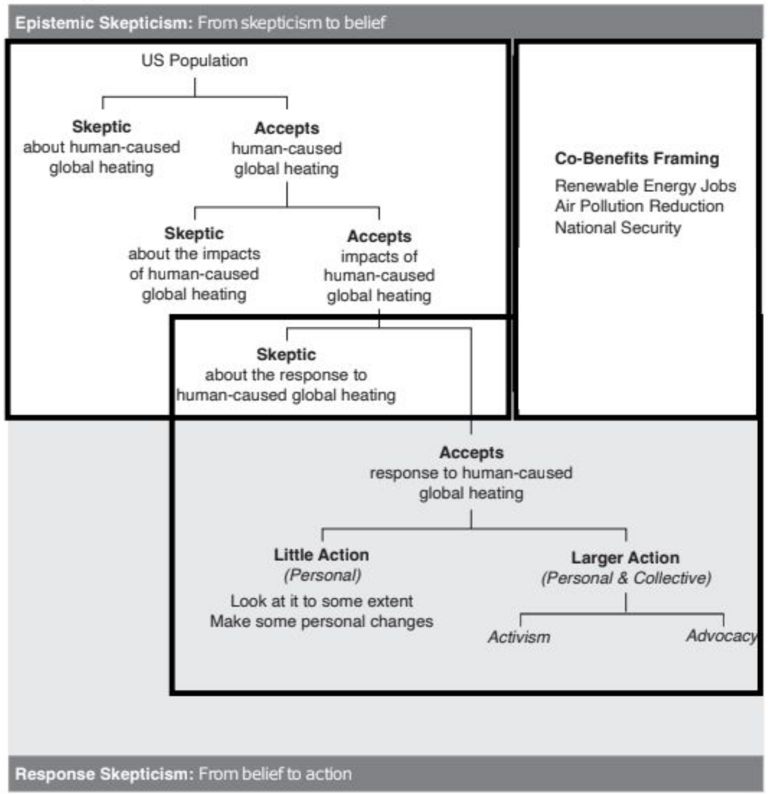
https://climatecommunication.yale.edu/publications/experience-with-global-warming-is-changing-peoples-minds-about-it/

Three ways to get a broader social mobilization:

Shift the skeptics to belief

Shift the believers to action

Leverage co-benefits to get policy support

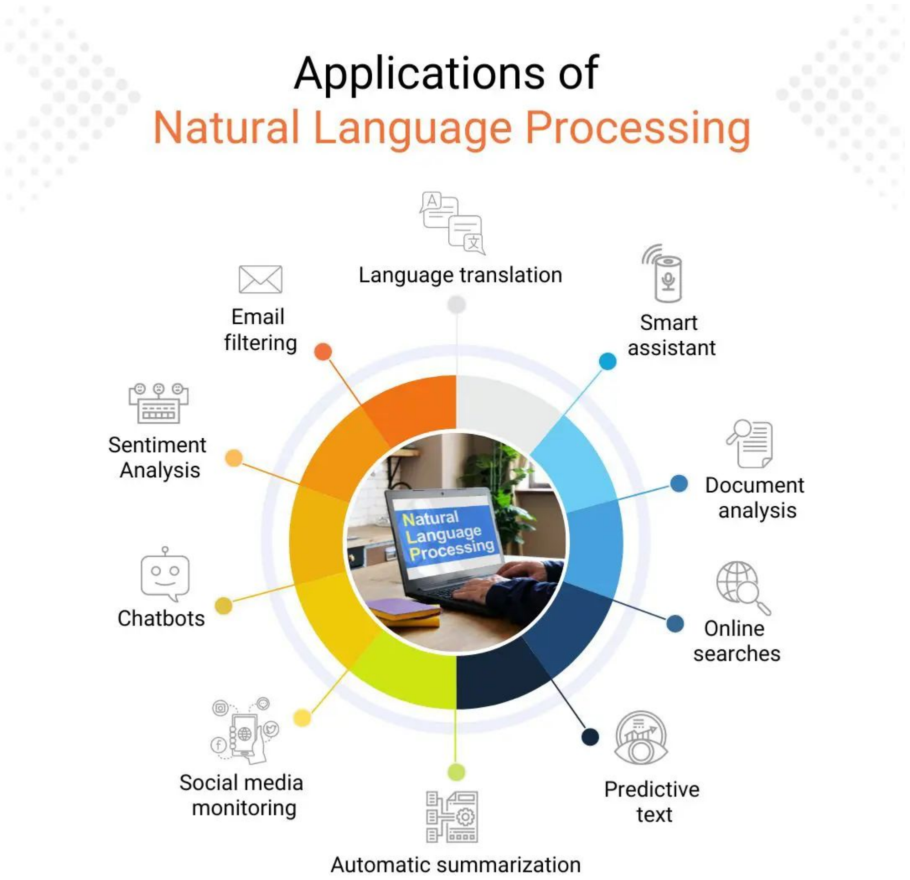


https://aronclimatecrisis.net/resources/

Natural Language Processing

NLP requires building algorithms that can make sense of text.

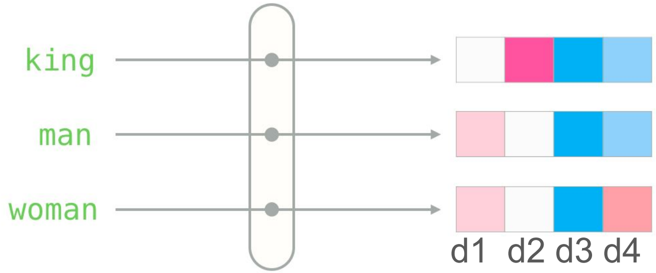
NLP tasks can be incredibly challenging due to the diverse ways in which people use language and how language relates to the real world.

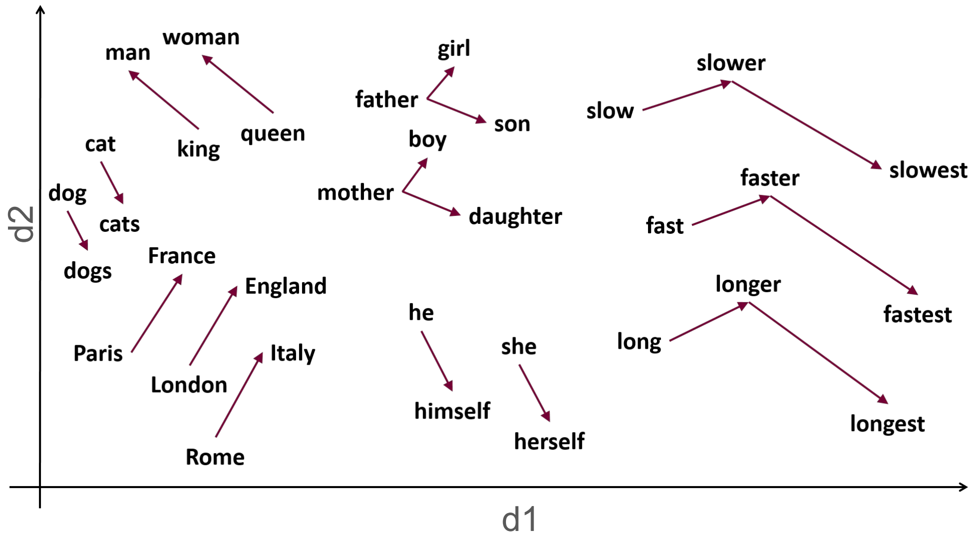


https://datasciencedojo.com/blog/natural-language-processing-applications/

Natural Language Processing

Requirement: Represent meaning as a vector of numbers





Natural Language Processing

Requirement: Represent meaning as a vector of numbers

Simplest approach = represent words in terms of how often they co-occur with other words.

	Roses	are	red	Sky	is	blue
Roses	1	1	1	0	0	0
are	1	1	1	0	0	0
red	1	1	1	0	0	0
Sky	0	0	0	1	1	1
is	0	0	0	1	1	1
Blue	0	0	0	1	1	1

What are the limitations of this?

Modern Approach

BERT: “Bidirectional Encoder Representations from **Transformers**”

Use a “Large Language Model” (LLM)

Attention Is All You Need

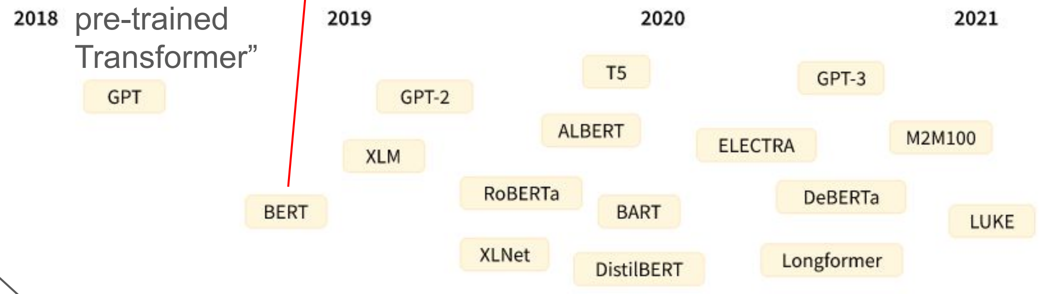
Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.0 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature.

Here are some reference points in the (short) history of Transformer models:

“Generative pre-trained Transformer”



The Transformer architecture was introduced in June 2017. The focus of the original research was on translation tasks. This was followed by the introduction of several influential models, including:

“Foundation Models”

As we saw previously, we can use transfer learning/fine-tuning on a pre-trained model to solve tasks where data is limited.

Foundation models take this idea to the extreme.

— WHAT IS A FOUNDATION MODEL?

In recent years, a new successful paradigm for building AI systems has emerged: Train one model on a huge amount of data and adapt it to many applications. We call such a model a foundation model.

— WHY DO WE CARE?

Foundation models (e.g., GPT-3) have demonstrated impressive behavior, but can fail unexpectedly, harbor biases, and are poorly understood. Nonetheless, they are being deployed at scale.

Our Mission

The Center for Research on Foundation Models (CRFM) is an interdisciplinary initiative born out of the Stanford Institute for Human-Centered Artificial Intelligence (HAI) that aims to make fundamental advances in the study, development, and deployment of foundation models.

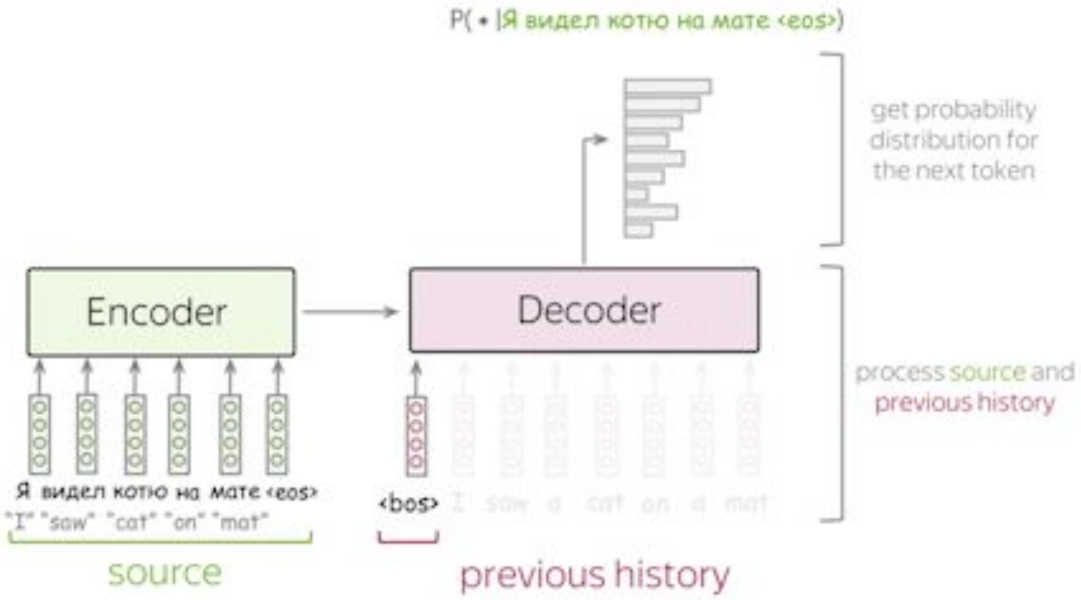
We are an interdisciplinary group of faculty, students, post-docs, and researchers spanning 10+ departments who have a shared interest in studying and building responsible foundation models.

> **CRFM has the following thrusts:**
>
> - **Research.** We will conduct interdisciplinary research that lays the groundwork of how foundation models should be built to make them more efficient, robust, interpretable, multimodal, and ethically sound.
> - **Artifacts.** We will train and release foundation models, code, tools, and also ensure that the full training pipeline is reproducible and scientifically rigorous.
> - **Community.** We will invite universities, companies, and non-profits to convene and work together to develop a set of professional norms for how to responsibly train and deploy foundation models.

Essentially, foundation models *learn good representations*.

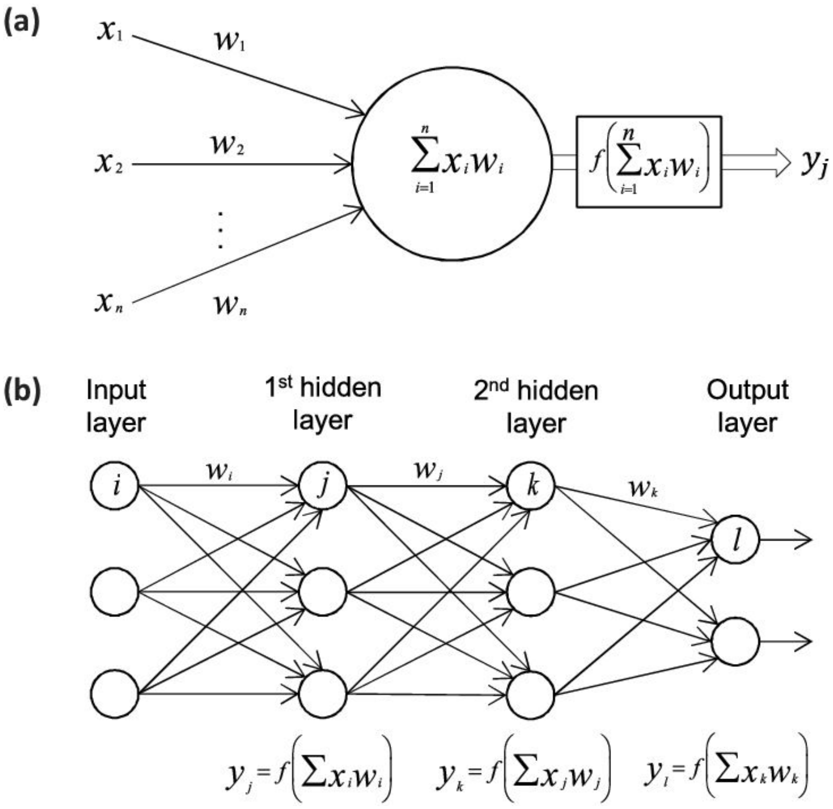
Architecture of a Large Language Model

Many language tasks are “sequence to sequence” problems that can be solved with an encoder and decoder. The encoder and decoder are each artificial neural networks

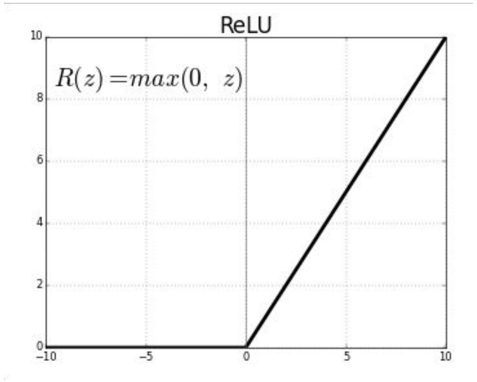


https://lena-voita.github.io/nlp_course/seq2seq_and_attention.html

Neural networks

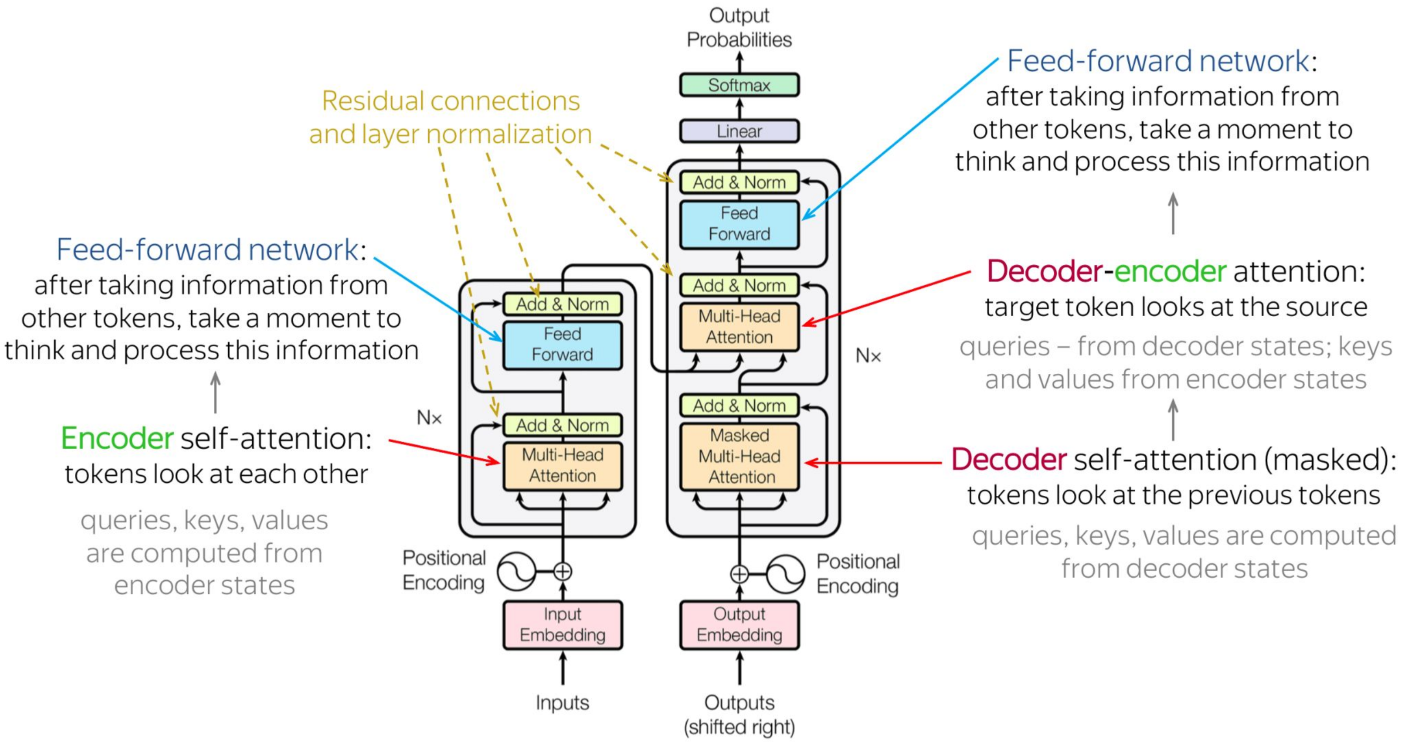


Basic or “vanilla” networks multiple weights by node activity, sum these values, and rectify the sum.



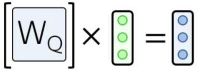
Vieira et al.

Transformer architecture



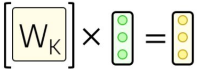
Transformer architecture

Each vector receives three representations ("roles")



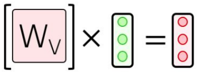
Query: vector **from** which the attention is looking

"Hey there, do you have this information?"



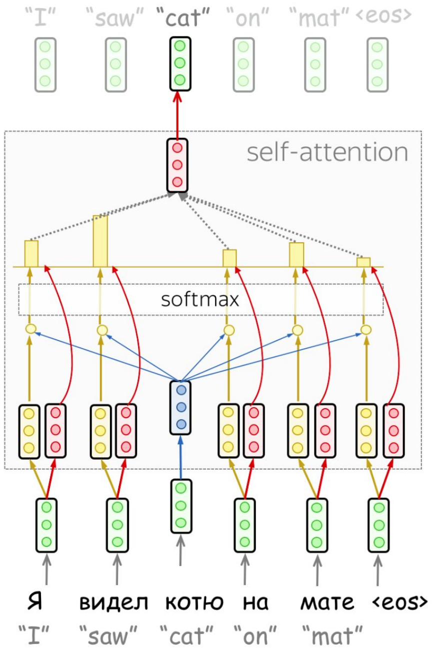
Key: vector **at** which the query looks to compute weights

"Hi, I have this information – give me a large weight!"



Value: their weighted sum is attention output

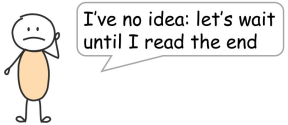
"Here's the information I have!"



Key insight: combine information across words. This is known as "self-attention".

I arrived at the bank after crossing the ... ...street? ...river?

What does **bank** mean in this sentence?



RNNs

O(N) steps to process a sentence with length N

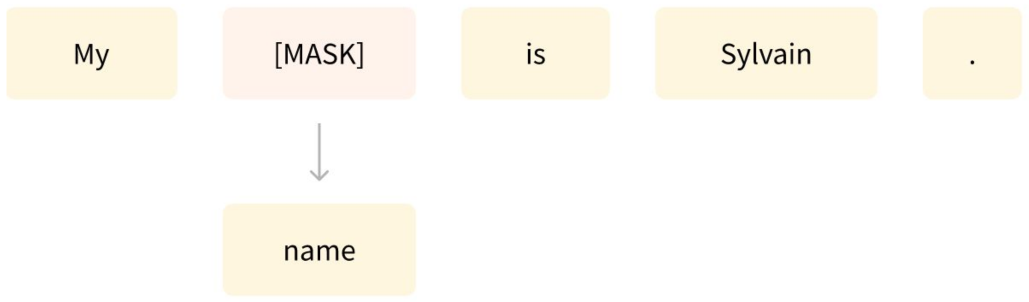
Transformer

Constant number of steps to process any sentence

LLMs can be trained on many different tasks

For example: language translation or next word prediction (ChatGPT)

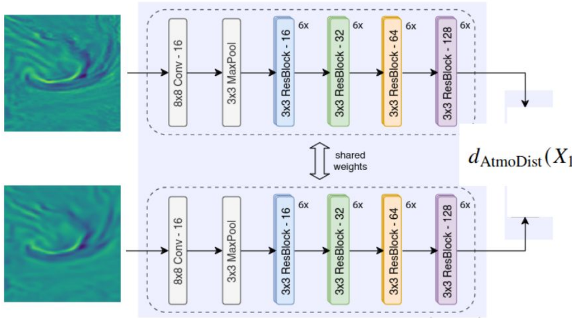
BERT is trained with a “masking” task: predict hidden word.



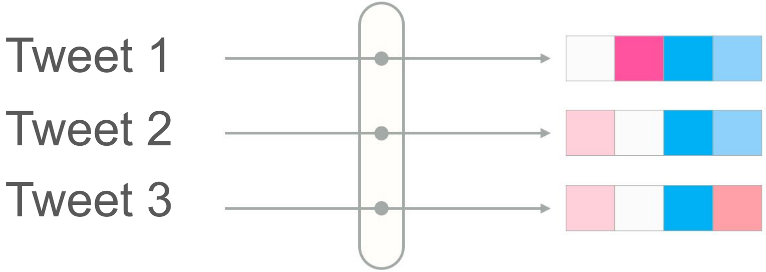
The LLM gives us a new representation

This is also known as the “embedding space”

	Roses	are	red	Sky	is	blue
Roses	1	1	1	0	0	0
are	1	1	1	0	0	0
red	1	1	1	0	0	0
Sky	0	0	0	1	1	1
is	0	0	0	1	1	1
Blue	0	0	0	1	1	1

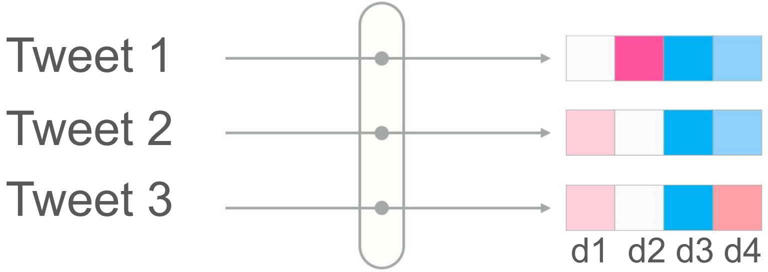


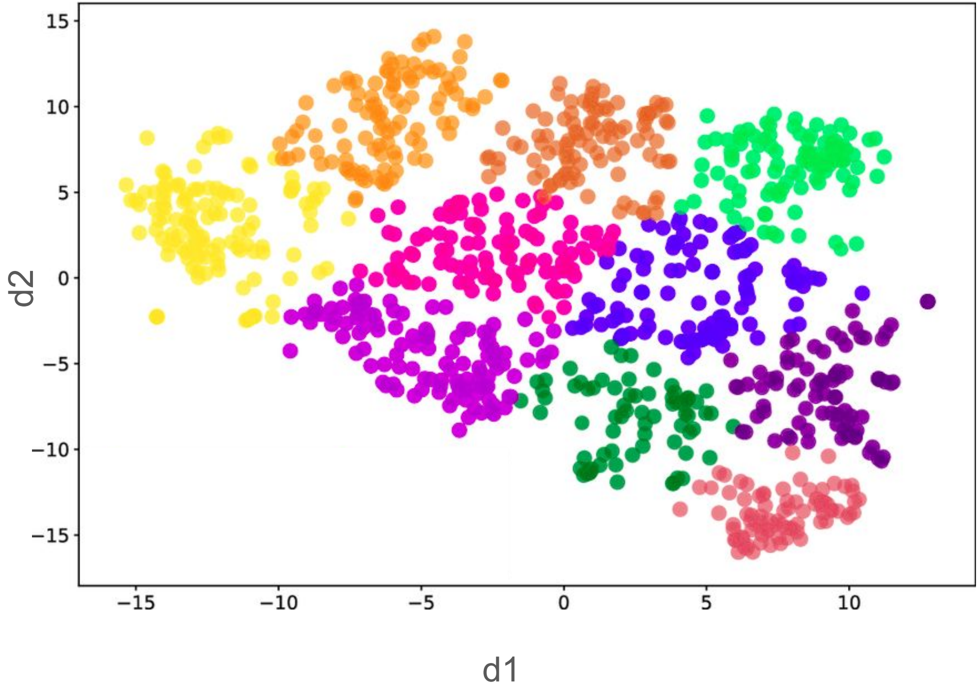
$$d_{\text{AtmoDist}}(X_1, X_2) = \frac{1}{N}\left\|\mathcal{F}(X_1) - \mathcal{F}(X_2)\right\|^2$$



The LLM gives us a new representation

We can do things like in that space, like unsupervised clustering





The output of the clustering gives us a new “topic” based representation we can use for downstream tasks

For your reading and your homework:

In machine learning settings where hyperparameters need to be set, data is typically divided into three subsets:

Training - data you actually pass to the algorithm that it uses to update weights

Validation - data you use to test the performance of models with different hyperparameters

Test - data you use to evaluate your model once you have decide on the hyperparameters