Interpretable Neural Networks for Learning New Science

Elizabeth A. Barnes, Associate Professor, Dept. of Atmospheric Science, CSU

Collaborators for slides in this talk

Benjamin Toms, PhD student, CSU

Imme Ebert-Uphoff, Research Faculty, CSU

Patrick Keys, Research Scientist, CSU

SiPN2 Seminar
July 29, 2020

Machine learning for science

data

prediction

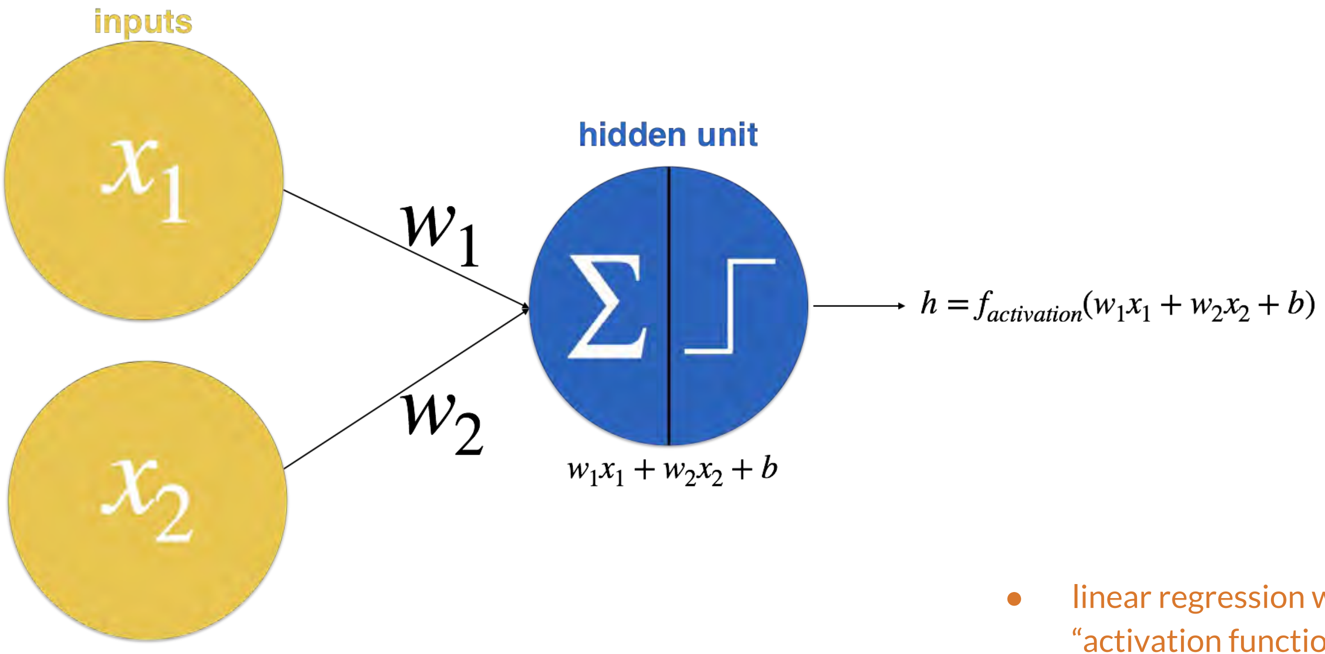
Machine learning for science

data →  → prediction

Not a black box!

Visualization tools are a *game changer* for using machine learning methods for science.

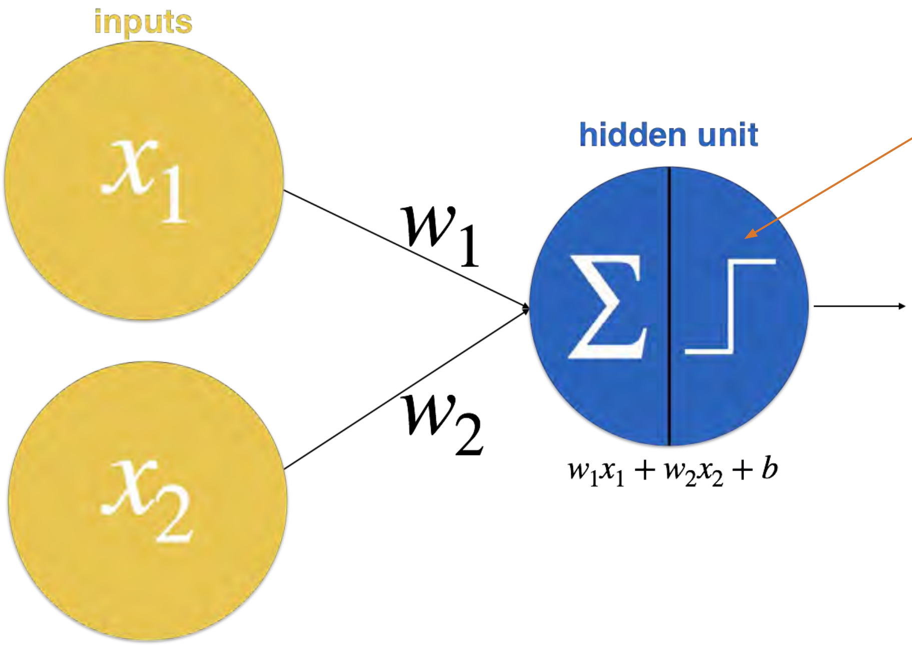
Artificial Neural Networks [ANN]



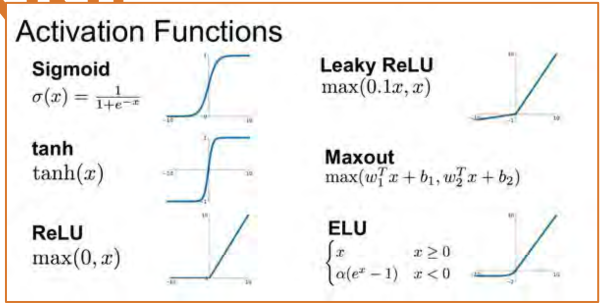
e.g. gridded sea surface temperatures

- linear regression with non-linear mapping by an "activation function"
- training of the network is merely determining the weights "w" and bias/offset "b"

Artificial Neural Networks [ANN]



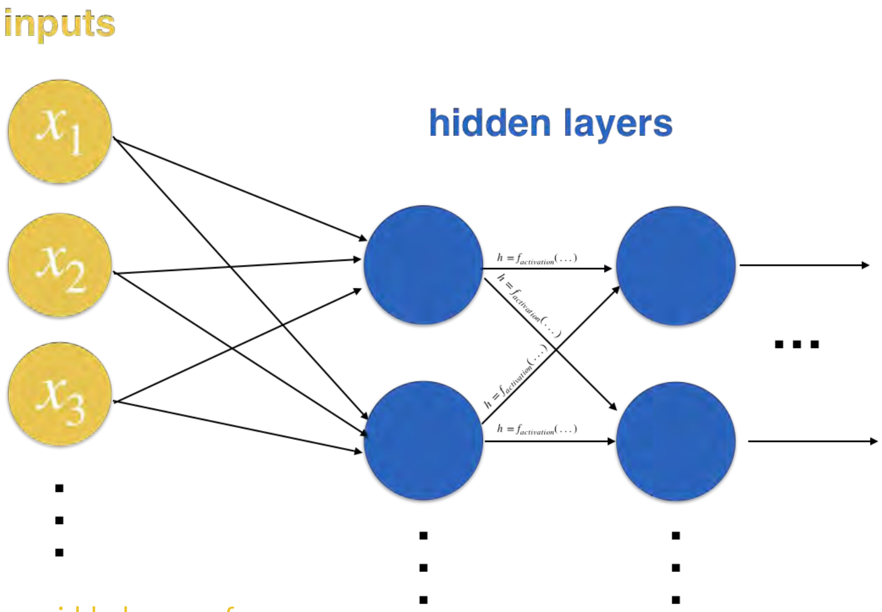
$$h = f_{activation}(w_1x_1 + w_2x_2 + b)$$



e.g. gridded sea surface temperatures

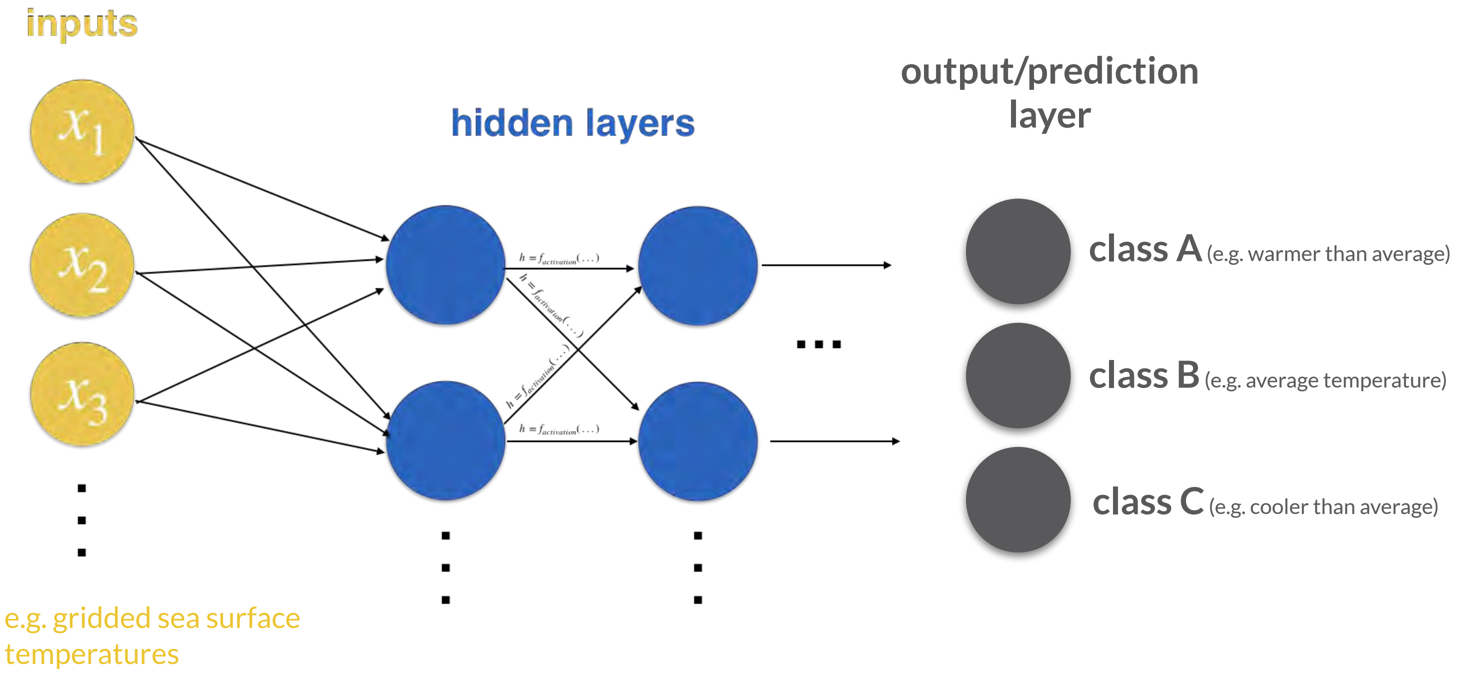
- linear regression with non-linear mapping by an "activation function"
- training of the network is merely determining the weights "w" and bias/offset "b"

Artificial Neural Networks [ANN]

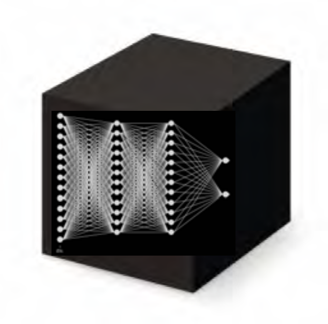


e.g. gridded sea surface temperatures

Artificial Neural Networks [ANN]

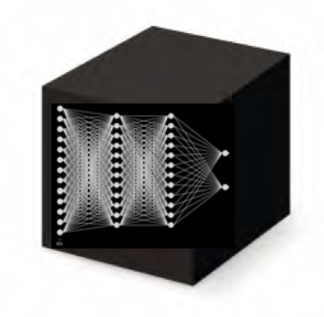


Artificial Neural Networks [ANN]

data →  → prediction

- Complexity and nonlinearities of the ANN allow it to learn many different pathways of predictable behaviour
- Once trained, you have an array of weights and biases which can be used for prediction on new data

Artificial Neural Networks [ANN]

data →  → prediction

- Complexity and nonlinearities of the ANN allow it to learn many different pathways of predictable behaviour
- Once trained, you have an array of weights and biases which can be used for prediction on new data
- But, how did the network make its prediction? What did it learn?

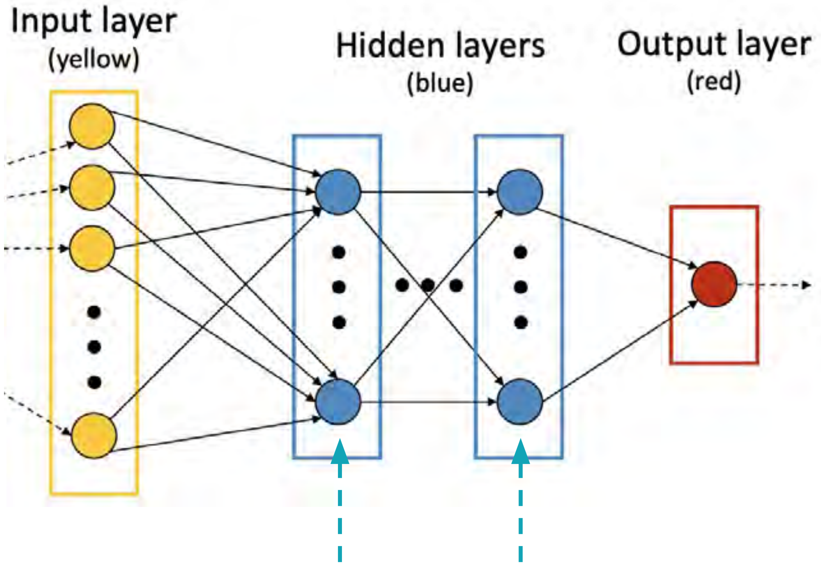
What to expect from ANN visualization

Not a perfect view, but better than the “black box”.

Two types of visualization tools

Type A: Feature Visualization

Philosophy: Seek to understand all internal components of ANN.

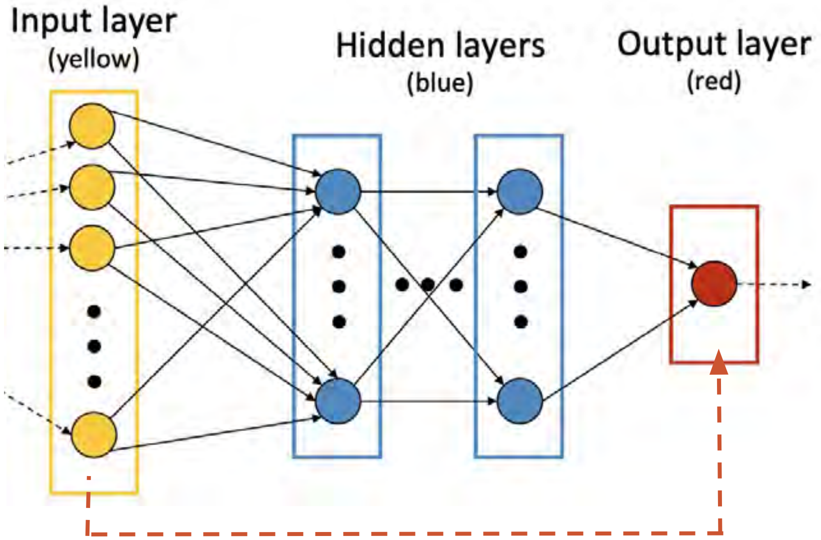


Seek to understand the **meaning of all intermediate (blue) nodes.**

Two types of visualization tools

Type B: Attribution / Explaining Decisions

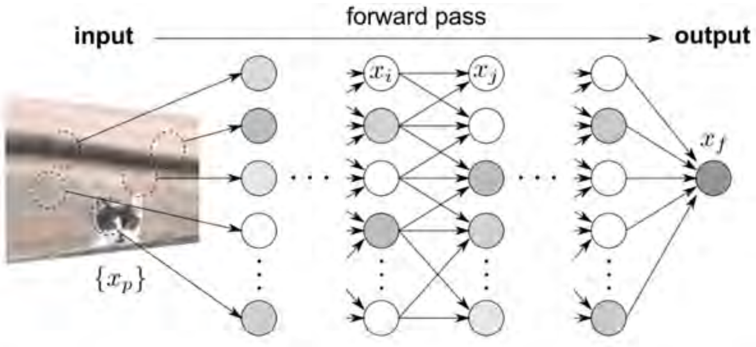
Philosophy: Understand the ANN's overall decision making for specific input.



Seek to understand the meaning of the entire algorithm - for a specific input.
Do NOT worry about meaning of intermediate (blue) nodes.

A visualization tool: Layerwise Relevance Propagation

Prediction
of 1 sample



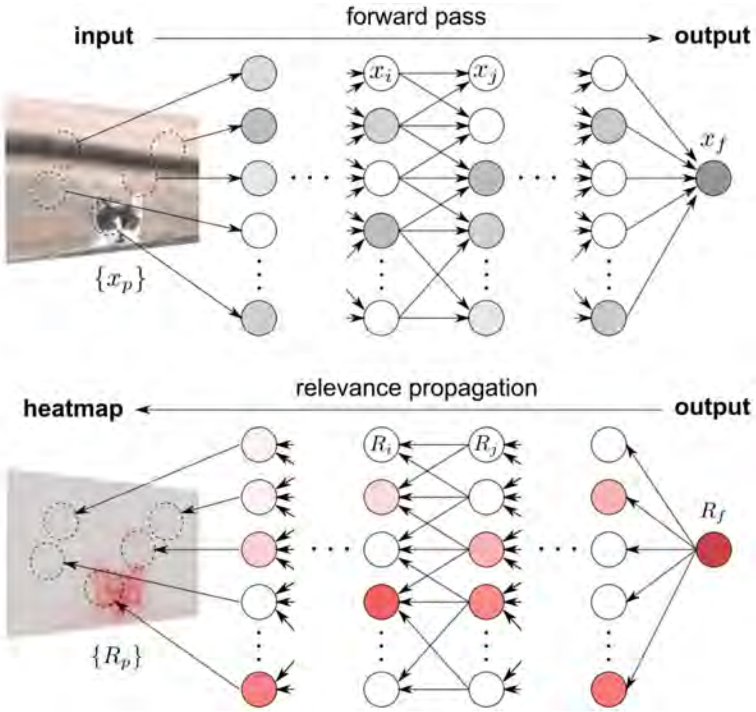
Pr(cat)

Montavon et al. (2017), Pattern Recognition; Montavon et al. (2018), Digital Signal Processing

A visualization tool: Layerwise Relevance Propagation

Prediction

of 1 sample



Pr(cat)

LRP

of 1 sample

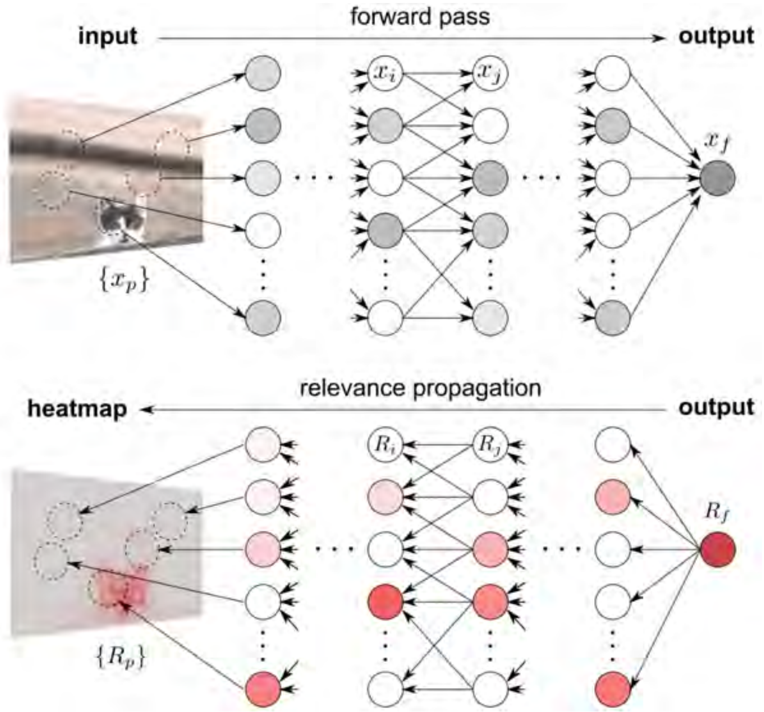
Pr(cat)

Montavon et al. (2017), Pattern Recognition; Montavon et al. (2018), Digital Signal Processing

A visualization tool: Layerwise Relevance Propagation

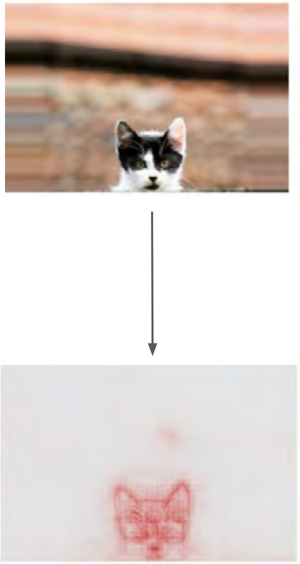
Prediction
of 1 sample

LRP
of 1 sample



Pr(cat)

Pr(cat)



where the network looked to determine it was a "cat"

Montavon et al. (2017), Pattern Recognition; Montavon et al. (2018), Digital Signal Processing

Example use of LRP

Task: Decide whether there is a horse in a given image.

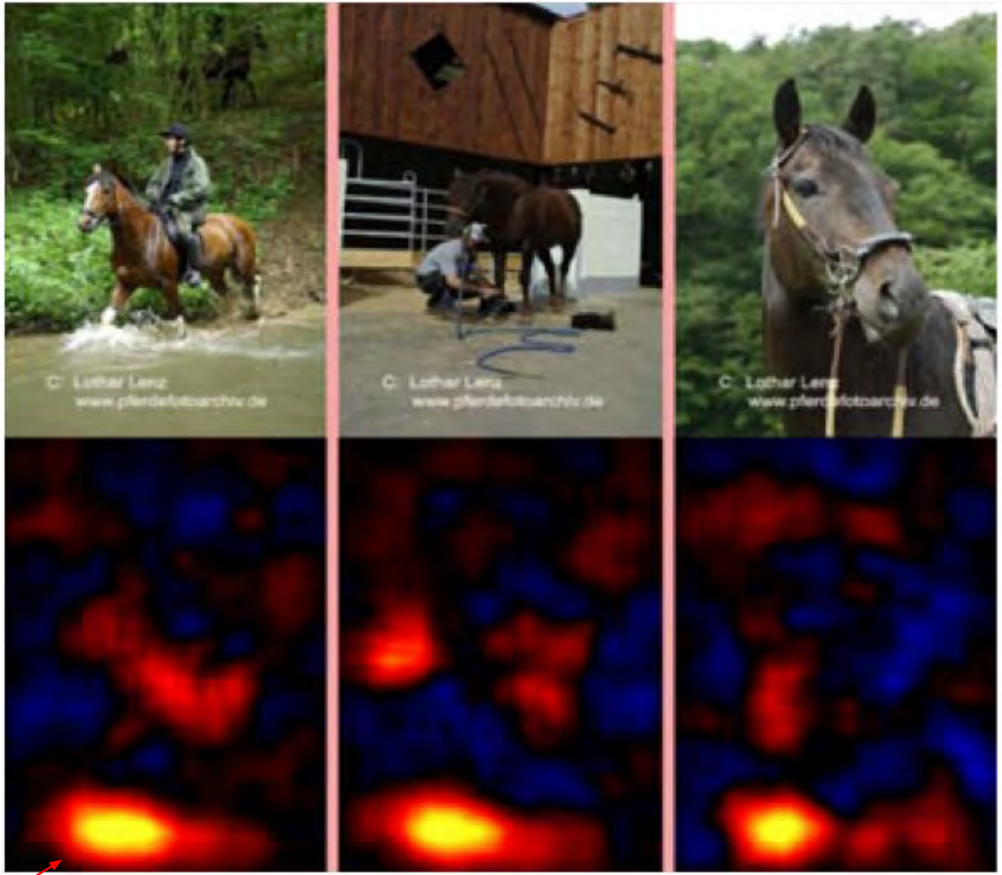
Decision making strategy: use visualization tools to determine the strategy the network used to make a decision

Lapuschkin et al. (2019)

Example use of LRP

Task: Decide whether there is a horse in a given image.

Decision making strategy: use visualization tools to determine the strategy the network used to make a decision



regions relevant to the network's decision

Lapuschkin et al. (2019)

What does this mean for earth system prediction research?

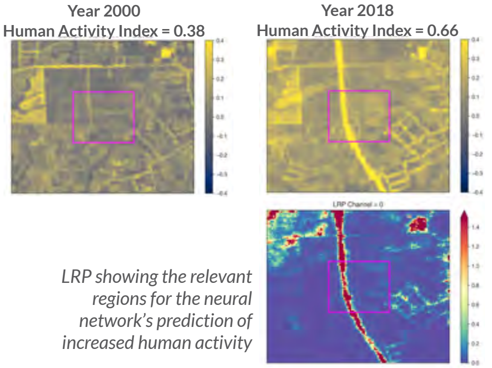
1. Identifying problematic strategies (i.e. right answer for the wrong reasons)
2. Designing the machine learning methodology
3. Building trust

What does this mean for earth system prediction research?

1. Identifying problematic strategies (i.e. right answer for the wrong reasons)

2. Designing the machine learning methodology

3. Building trust



Landsat imagery

LRP showing the relevant regions for the neural network's prediction of increased human activity

Keys and Barnes (2020; in prep)

What does this mean for earth system prediction research?

1. Identifying problematic strategies (i.e. right answer for the wrong reasons)
2. Designing the machine learning methodology
3. Building trust
4. **Discovering new science!**
    - **When** our machine learning method is capable of making an accurate prediction we can explore **why**

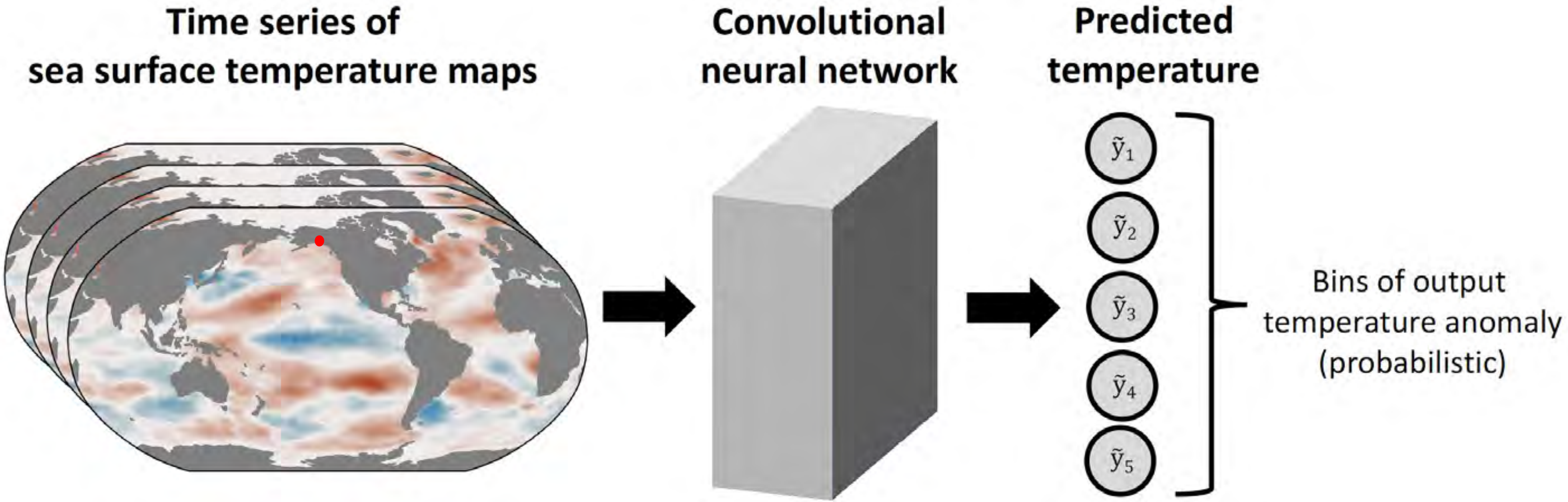
Science Applications

1. Multi-year prediction
2. Subseasonal-to-seasonal prediction
3. Indicator patterns of forced change
4. Eddy-mean flow interactions
5. Human impacts on the land surface from Landsat imagery

Science Applications

1. **Multi-year prediction**
2. Subseasonal-to-seasonal prediction
3. Indicator patterns of forced change
4. Eddy-mean flow interactions
5. Human impacts on the land surface from Landsat imagery

Multi-year prediction network set-up

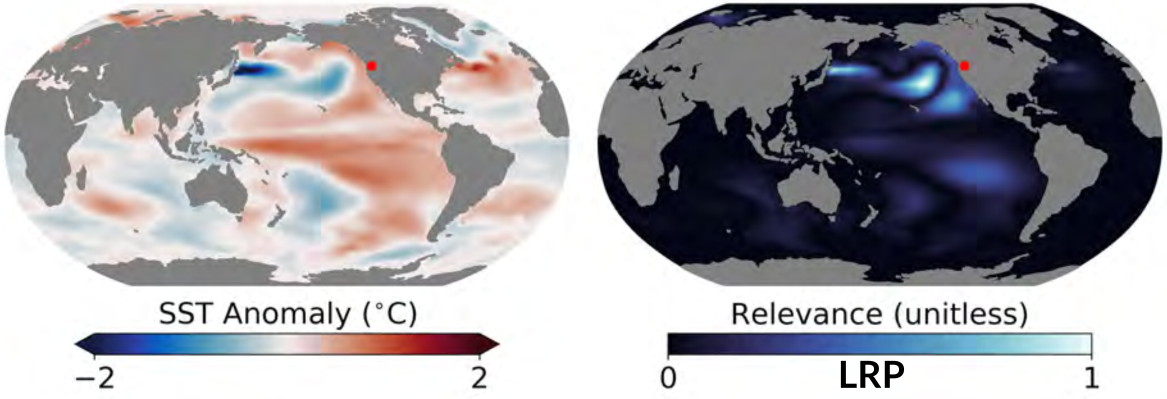


Predicting 5-year average surface temperature at one grid point
Applied to 1200 years of CESM2 control simulation
Toms et al. (2020; in prep)

Examples of neural network-driven predictions

- Neural network + LRP can be used to identify patterns of earth-system variability that lend predictability
- Here, we **predict 5-year average surface temperature** using past maps of sea-surface temperature
- Each prediction uses spatially unique information, although dominant patterns emerge

example accurate prediction

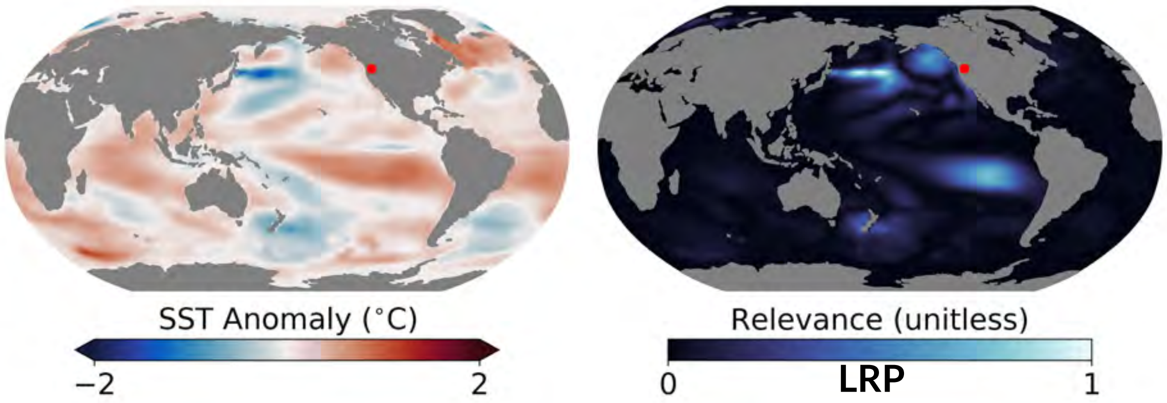


Predicting 5-year average surface temperature at one grid point
Applied to 1200 years of CESM2 control simulation
Toms et al. (2020; in prep)

Examples of neural network-driven predictions

- Neural network + LRP can be used to identify patterns of earth-system variability that lend predictability
- Here, we **predict 5-year average surface temperature** using past maps of sea-surface temperature
- Each prediction uses spatially unique information, although dominant patterns do exist

example accurate prediction

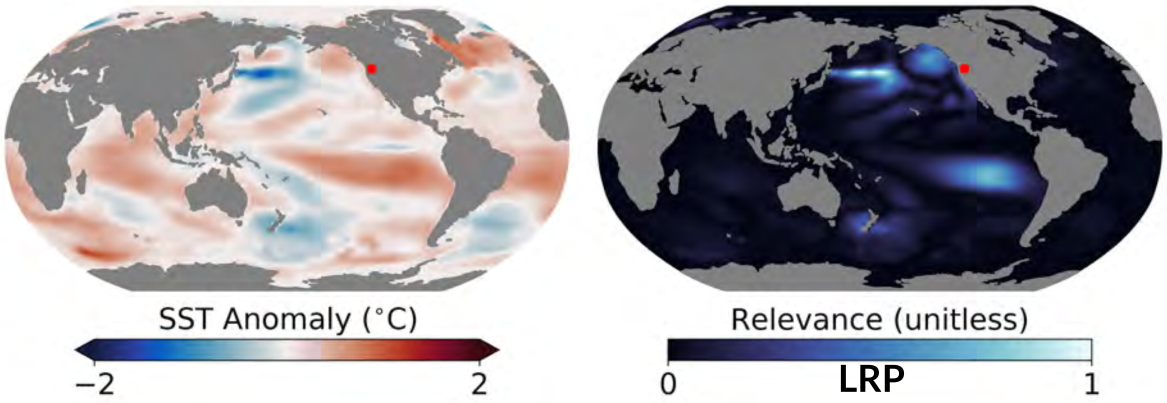


Predicting 5-year average surface temperature at one grid point
Applied to 1200 years of CESM2 control simulation
Toms et al. (2020; in prep)

Examples of neural network-driven predictions

- Neural network + LRP can be used to identify patterns of earth-system variability that lend predictability
- Here, we **predict 5-year average surface temperature** using past maps of sea-surface temperature
- Each prediction uses spatially unique information, although dominant patterns do exist

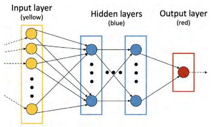
example accurate prediction



For us, the science is not the making of a multi-year prediction - it is **identifying predictable patterns/regimes** of the earth system

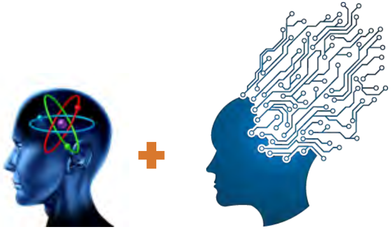
Predicting 5-year average surface temperature at one grid point
Applied to 1200 years of CESM2 control simulation
Toms et al. (2020; in prep)

Wrap-up



- The most basic of neural networks can be viewed as nonlinear regression - **climate scientists are well-equipped** to think about this architecture

- Artificial neural networks are **no longer black boxes -** tools exist to help **visualize their decisions**. This is a **game changer** for their use in geoscience research.



- ANNs can be used for more than just prediction. The **science can be what the network learns**, rather than the prediction. **Get creative** combining your science with these tools!

Elizabeth A. Barnes
eabarnes@rams.colostate.edu,
Twitter @atmosbarnes

References

- **Introduction of LRP to the geosciences:**
  Toms, Benjamin A., Elizabeth A. Barnes, and Imme Ebert-Uphoff: Physically interpretable neural networks for the geosciences: Applications to earth system variability, *JAMES*, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019MS002002.
- **Use of LRP for identifying patterns of climate change:**
  Barnes, Elizabeth A., Benjamin Toms, James Hurrell, Imme Ebert-Uphoff, Chuck Anderson and David Anderson: Indicator patterns of forced change learned by an artificial neural network, JAMES, under review, preprint available at http://arxiv.org/abs/2005.12322.
- **Use of LRP for identifying MJO variability:**
  Toms, B., K. Kashinath, Prabhat, and D. Yang (2020), Testing the Reliability of Interpretable Neural Networks in Geoscience Using the Madden-Julian Oscillation, Submitted to Geophysical Model Development (GMD), Preprint available: https://arxiv.org/abs/1902.04621.
- Ebert-Uphoff, I., & Hilburn, K. A. (2020). Evaluation, Tuning and Interpretation of Neural Networks for Meteorological Applications. Submitted to Bulletin of the American Meteorological Society (in review). Preprint available: https://arxiv.org/abs/2005.03126
- Lapuschkin et al. “Unmasking Clever Hans Predictors and Assessing What Machines Really Learn.” Nature Communications, vol. 10, no. 1, Mar. 2019, p. 1096, doi:10.1038/s41467-019-08987-4.
- Ebert-Uphoff, Imme, Savini Samarasinghe, and Elizabeth A. Barnes: Thoughtfully Using Artificial Intelligence in Earth Science, EOS, 100, https://doi.org/10.1029/2019EO135235.

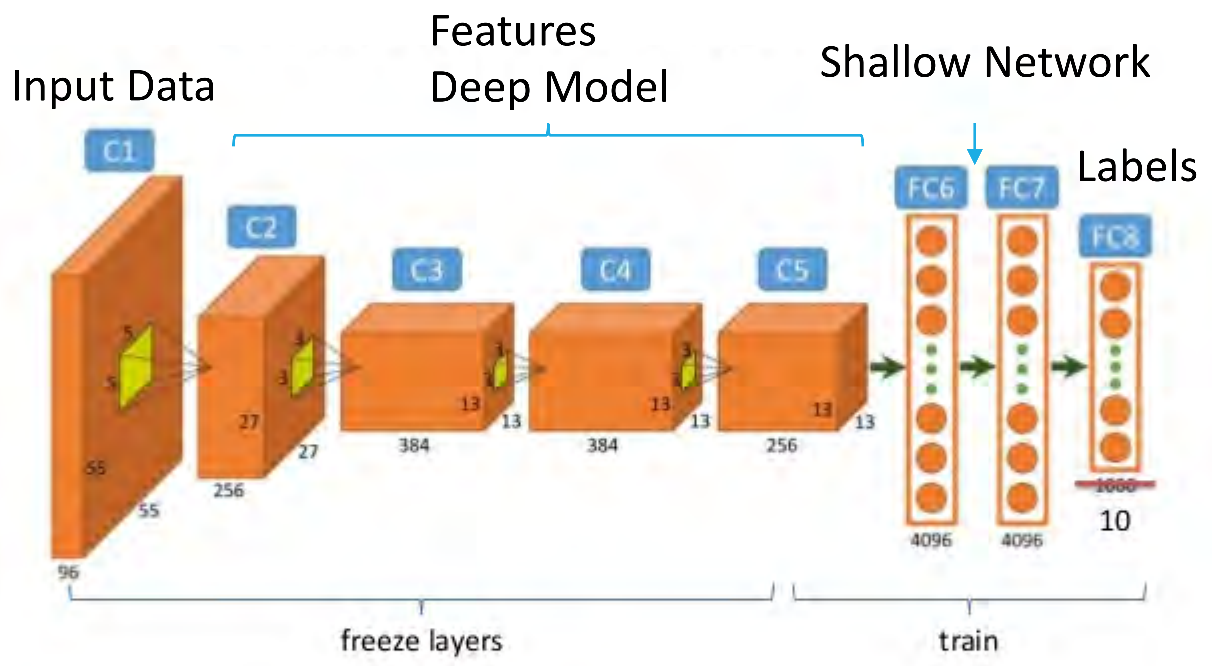
COLLEGE OR ARTS AND SCIENCES
DEPARTMENT OF GEOGRAPHY

Visual Analytics and Interactive Machine Learning for Geospatial Sciences and Cryospheric Research

MORTEZA KARIMZADEH, PH.D.
ASSISTANT PROFESSOR, GEOGRAPHY

ARCUS SIPN2 WEBINAR SERIES
JULY 29, 2020

Labeled Data and Pre-trained Models



Visual Analytics for Machine Learning

1. Real time social media analytics for situational awareness

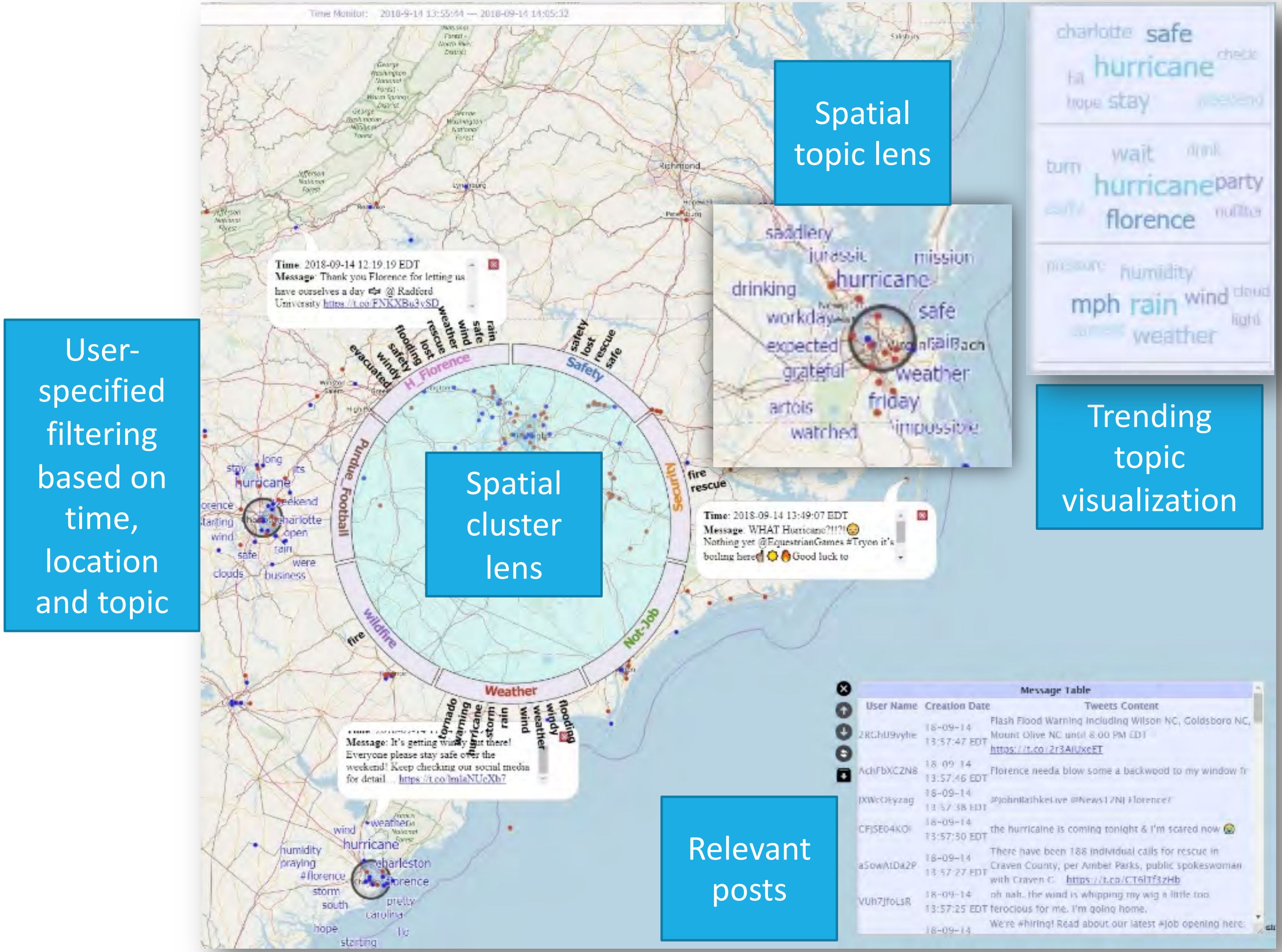
2. Spambot labeling and behavioral analysis

3. Upcoming NSF EarthCube project on Sea Ice mapping and classification

SMART

Situational awareness for first responders:

- Interactive interface
- Visualizations
- Topic modeling
- Advanced filtering
- Trends/anomalies



Harnessing Salient Information in Noisy Text

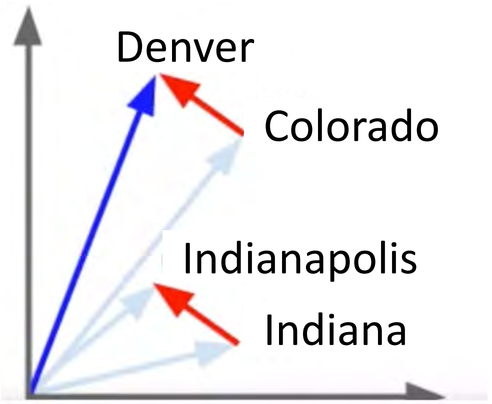
- How to reduce noise (irrelevant text).
  - Support <u>dynamic</u> phenomena.
    - *Spatial dimension.*
    - *Temporal dimension.*
    - *Semantic dimension.*
  - Support multilingual posts.

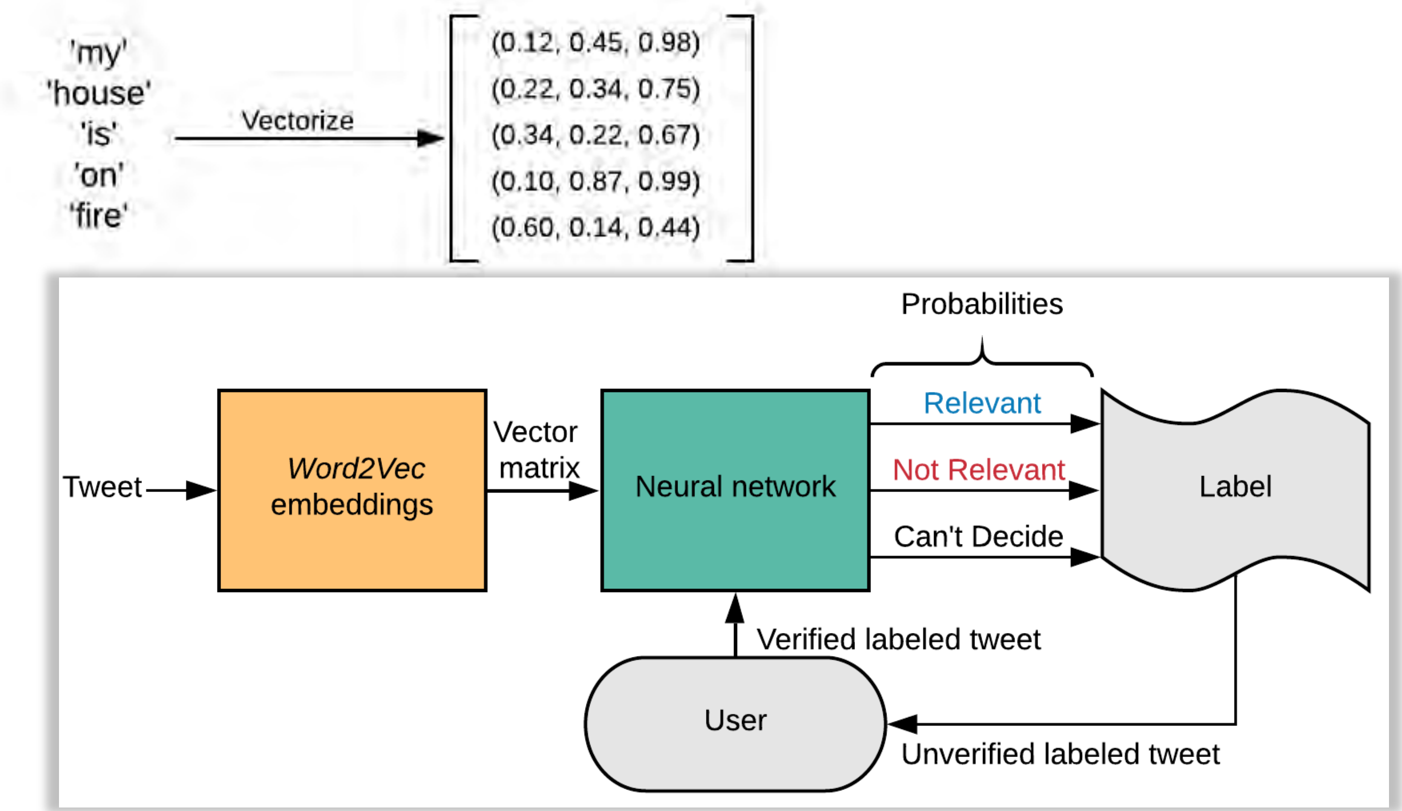
- Solution:
  - Interactively incorporate:
    - User knowledge
    - Linguistic context
      - *The entire apartment is burning down.* → ✓ Relevant
      - *Will Bernie feel the burn again?* → ✗ Not relevant

Human-in-the-loop Neural Networks

Transform words into a semantic space:

- Word2Vec : A model pre-trained on roughly 100 billion words, provides word embeddings (context of the target word), with each word represented as a 300-dimensional vector.



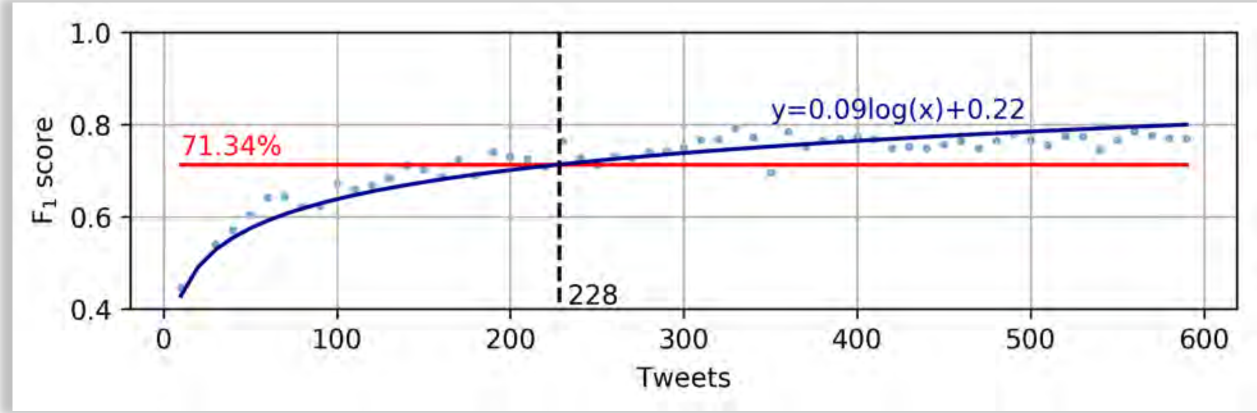


Evaluation

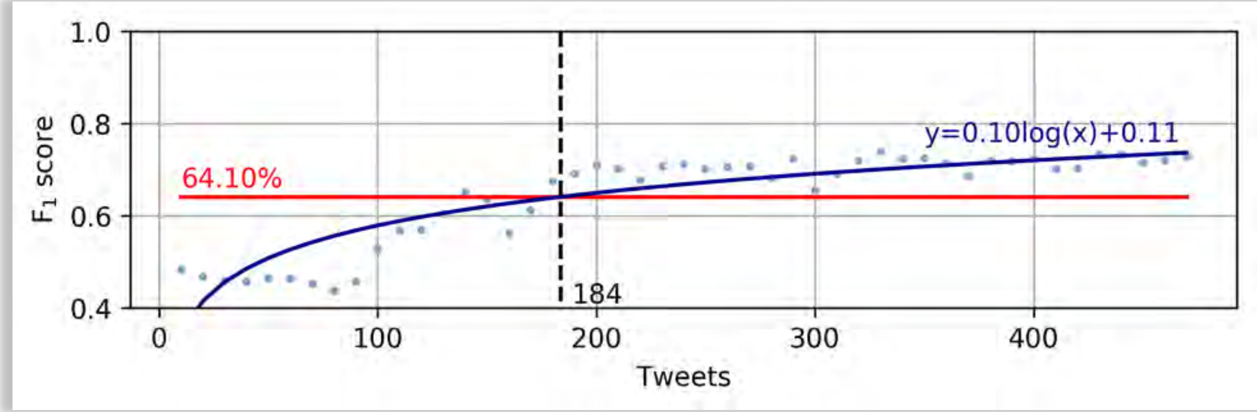
CrisisLexT26 datasets

- Trained iteratively with 10 tweets

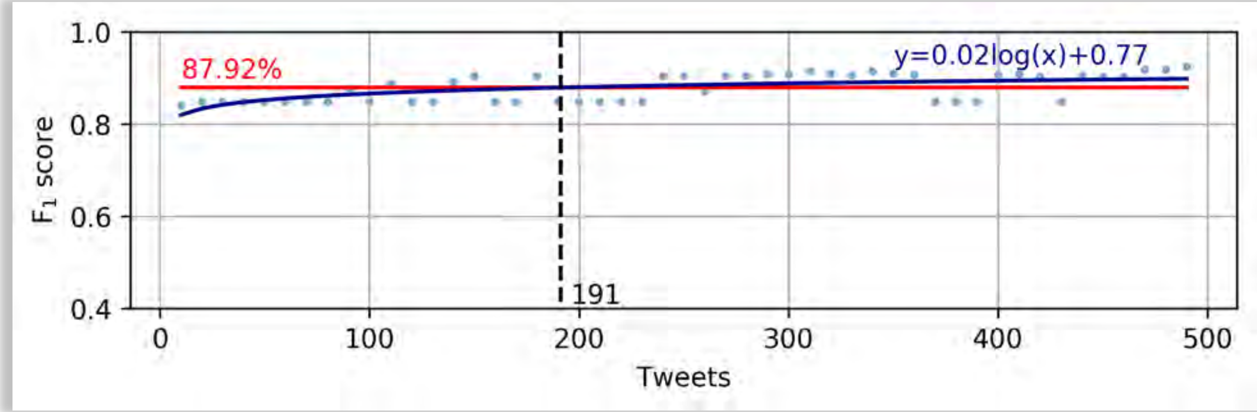
Model reaches its average F_1 score after approximately 200 tweets



2012 Colorado wildfires

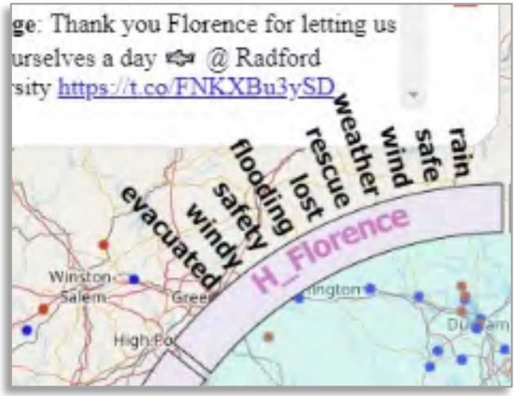


2013 Boston bombings



2013 NY train crash

Results after 20 Clicks…



The most relevant about weather events:

Message Table

User Name	Creation Date	Tweets Content	All	Relevant Probability	Not Relevant Probability	Can't Decide Probability
aPmcp5udJF	19-02-19 15:27:04 EDT	#DopplerGreg Storm Forecast: Snow, sleet, and rain across #NYC & #JerseyCity on Wednesday. ❄□□□ #NJWeather… https://t.co/VdGycsKzF7	Relevant	91.6%		
yRDYE0sr0I	19-02-19 16:51:00 EDT	@LeeGoldbergABC7 Another snow flop! Another rain/mix/slop!	Relevant	68.9%	25.4%	
cMVHsSkiul	19-02-19 12:16:25 EDT	#EWR is currently experiencing delays averaging 31 mins due to WEATHER / WIND #flightdelay https://t.co/seRNV1PL2a	Relevant	61.3%	38%	
ZjsB7cGGfF	19-02-19 16:30:23 EDT	Yay! Snow! ❄□□	Relevant	61.1%	26.9%	
2x0jzzhZfE	19-02-19 17:22:17 EDT	Yay! Snow!	Relevant	61.1%	26.9%	
Xd3z7jgZ0X	19-02-19 13:12:08 EDT	come out and play: a snow day anthem https://t.co/UcrwQim3QE	Relevant	60.2%	37.6%	
alP9lRmeYV	19-02-19 16:14:41 EDT	WINTER WEATHER ADVISORY The @NWSNewYorkNY has issued a winter weather advisory for the Cranford area. https://t.co/0EMgsNKkVo	Relevant	58.4%	38.4%	
QDpBXWBM8E	19-02-19 15:12:16 EDT	@SUNWAYHAWAII It's 36F now and snow tomorrow. Still wearing the double-lined furs https://t.co/cJAsgfexiP	Relevant	58%	40.7%	
DBmnwEldvz	19-02-19 15:58:19 EDT	#EWR is currently experiencing delays averaging 31 mins due to WEATHER / WIND #flightdelay https://t.co/seRNV1PL2a	Relevant	56.6%	42.8%	
jnXpLnOYuR	19-02-19 13:21:12 EDT	Super Snow Moon tonight. ❄ Biggest and brightest of 2019. No wonder I've been feeling "hinky" as I call it, today… https://t.co/RO7WyrW6S5	Relevant	55.5%	43.6%	

The least relevant about weather events:

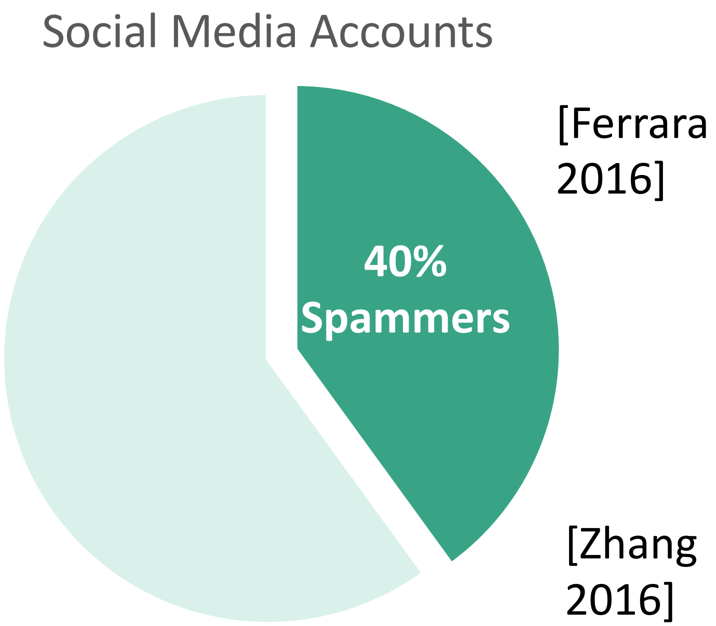
Message Table

User Name	Creation Date	Tweets Content	All	Relevant Probability	Not Relevant Probability	Can't Decide Probability
4VQSmQ9KHF	19-02-19 15:27:05 EDT	Can you recommend anyone for this #Java job in #NewYork, NY? Click the link in our bio to see it and more. Senior Risk Developer at Luxoft	Not Relevant		89%	
zPovuGeVs9	19-02-19 14:03:28 EDT	We're hiring in New York, NY! Click the link in our bio to apply to this job and more: Risk Specialist – NYC at PMA… https://t.co/13JDX5UksN	Not Relevant		83.7%	
D787V9zYep	19-02-19 16:36:34 EDT	Can you recommend anyone for this job? Manager, FCC Risk Assessment – https://t.co/GONYlXDOja #Legal #NewYork, NY	Not Relevant		82.9%	
zewhPrBvAJ	19-02-19 16:01:36 EDT	@HarlemXPancho It be too much. Like come ON NEW YORK! Just chill. If it ain't brick we have 30 feet of snow. Lol.	Not Relevant		78.9%	
oDDDpXPbQN	19-02-19 15:55:13 EDT	TMM🎬 BE VERY CAREFUL WHO YOU MAY JUDGE WHEN GOD SENT THEM TO HELP YOU...WARNING🚨 I PRAY OVER MY SELF ON DAILY TO G… https://t.co/My6HSJp6Wm	Not Relevant		78.8%	
ttt1RYEOgq	19-02-19 16:04:04 EDT	@katiecannon2 @mark_dow @MazzucatoM Of course there are. Though I think governments acting as "first risk takers" i… https://t.co/enTVddjuzy	Not Relevant	20.4%	78.1%	
So89zjKYPe	19-02-19 15:52:48 EDT	@flyaway_k @IcyVoteblue □THESES PEOPLE WILL LIE.RIGHT IN FRONT YOUR FACE.□IF YOU TOLD THEM SNOW IS WHITE.*OH NO ITS… https://t.co/vTizQ7u2lR	Not Relevant		77.2%	
Zg1fga8Zmw	19-02-19 15:36:17 EDT	Got my run in today. That wind was cold today ➾➾ almost didn't go today but in glad I pushed myself to hit the ro… https://t.co/ln4SH5L0Ca	Not Relevant		76%	
	19-02-19	Be. Stay. Think positive! It's □ and the weather is not				

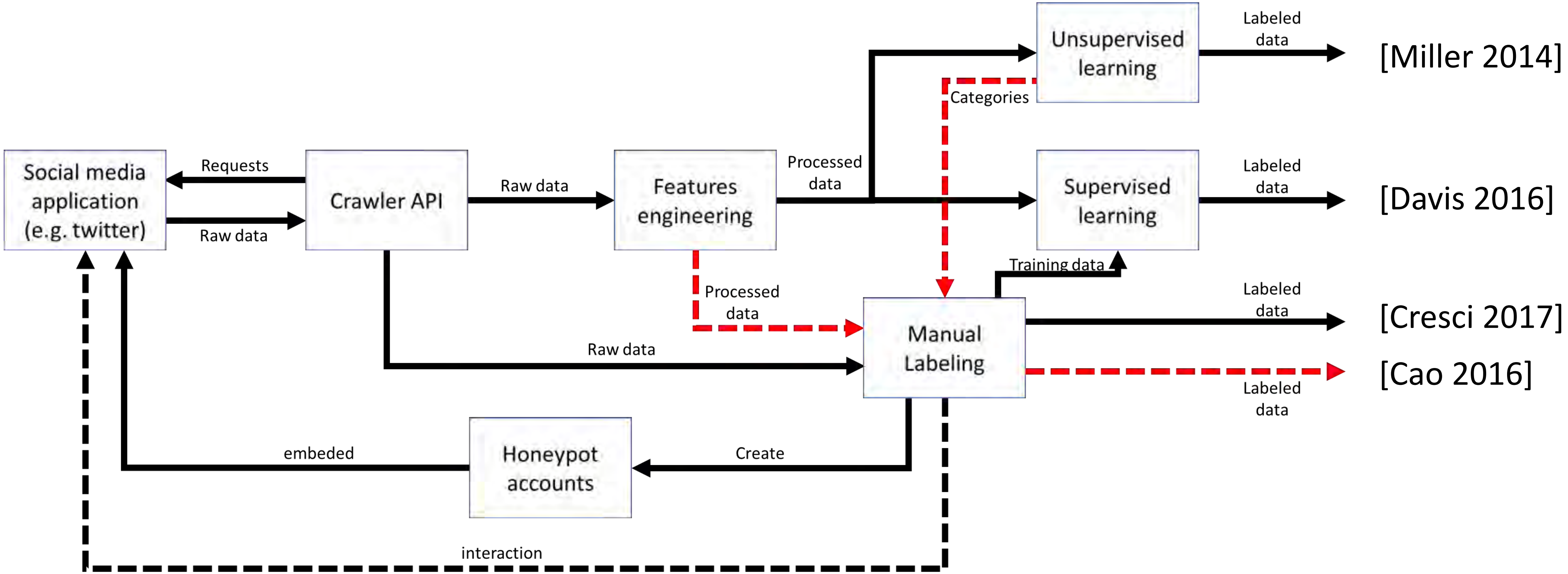
Social Spambot

A computer algorithm that automatically produces content and interacts with humans on social media, trying to emulate and possibly alter their behavior.

- Spread disinformation
- Manipulate public opinions
- Distribute unsolicited spam
- Propagate malicious links
- Steal personal information



Existing Automated and VA solutions

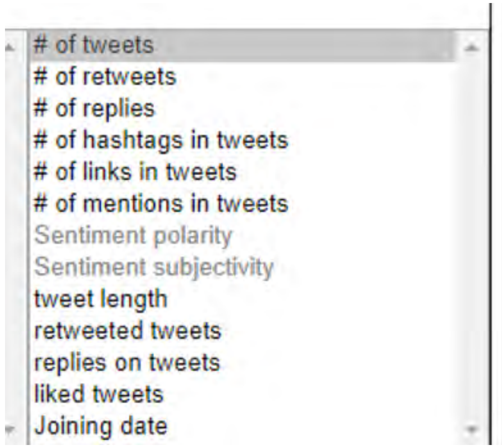


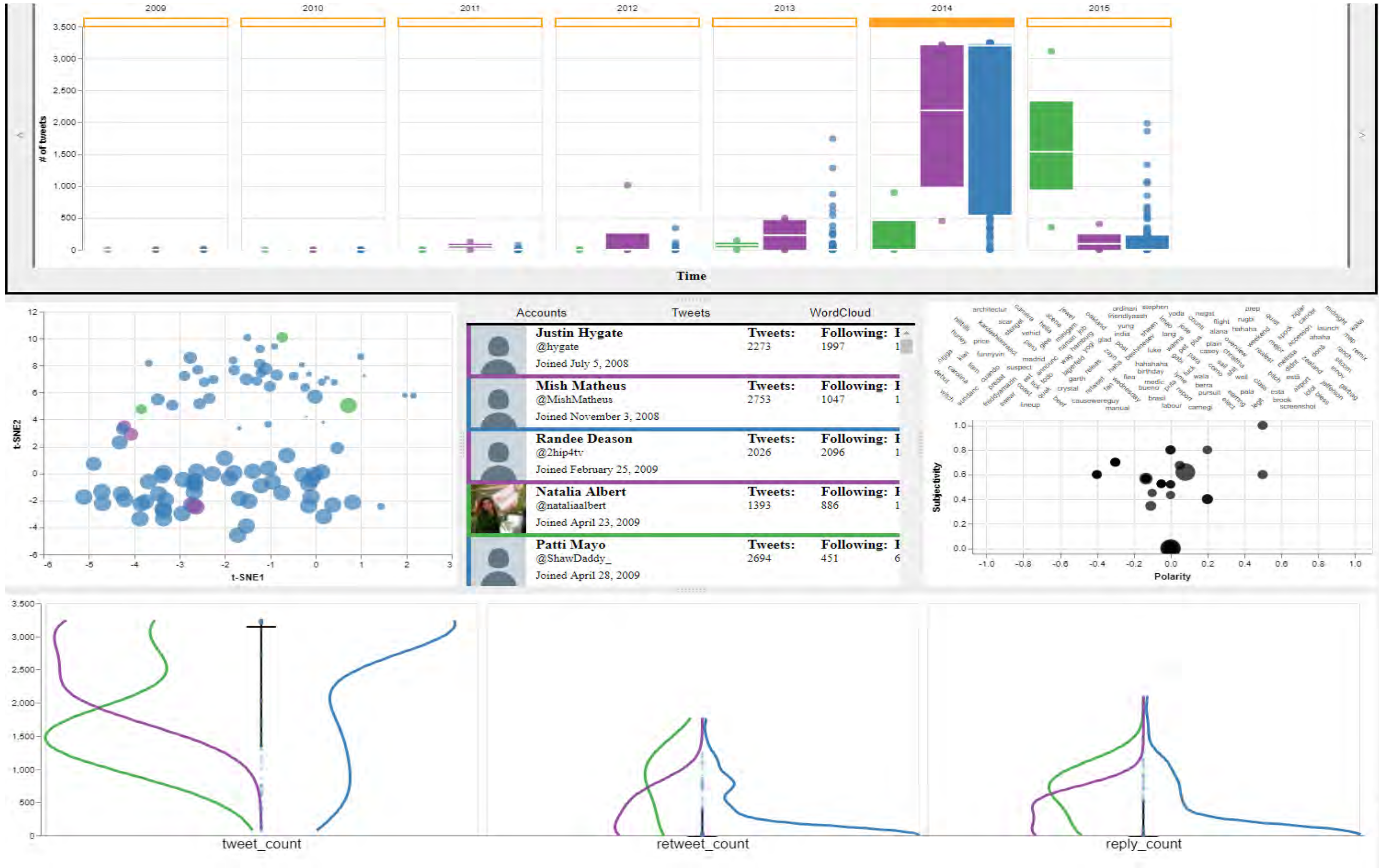
Issues

- Spambots with natural behavior at individual level➔ Harder to detect **spam groups/campaigns**
- Continually Changing Environment ➔ Effort to **maintain representative** training set

Visual Analytics for Social Spambot Labeling (VASSL)

- Output labels: Spambot or genuine
- Input:
  - Tweet Text
  - Metadata:





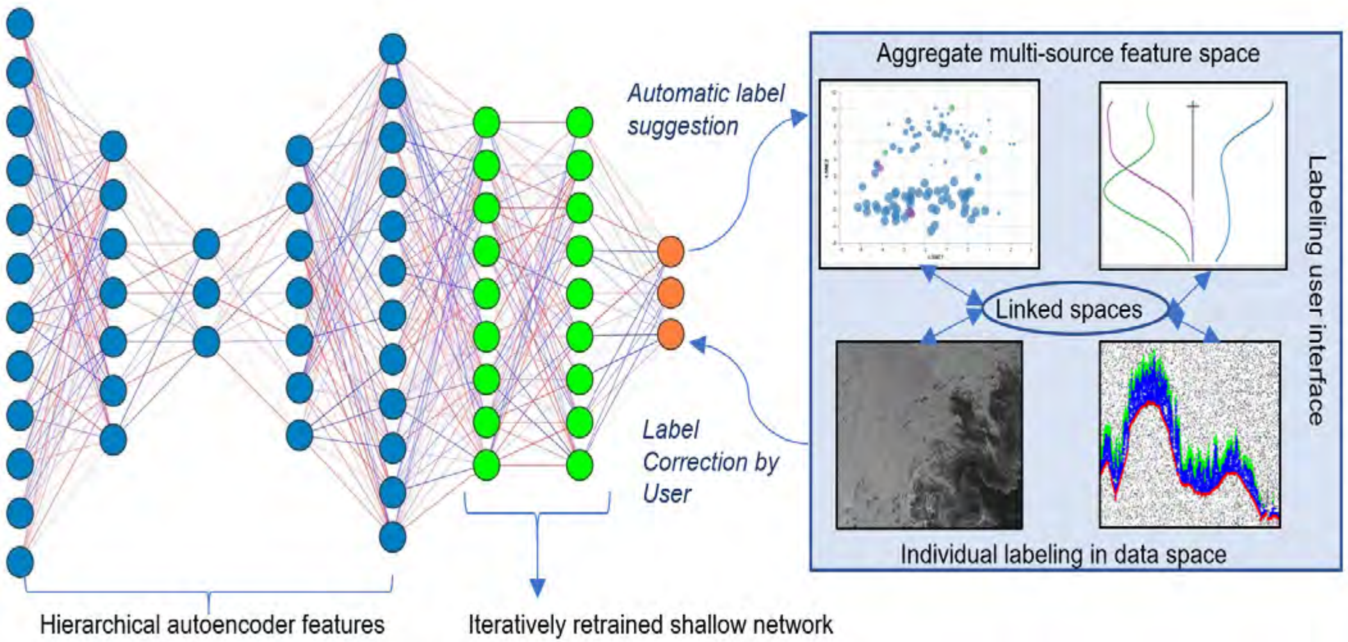
Khayat, M., Karimzadeh, M., Zhao, J., & Ebert, D. S. (2020). VASSL: A Visual Analytics Toolkit for Social Spambot Labeling. *IEEE Transactions on Visualization and Computer Graphics*.


2008
2009
2010
2011
2012
2013
2014
2015
# of tweets
Time
General control panel
Spambot
Genuine
Unlabeled
Selected
Reset Layout
Account Selector tools:
Select All
Select None
Select Inverse
Select Spambots
Select Genuine
Select Unlabeled
Selection rule: New Add Subtract
Auto Open Control Panel for the on Focus View
Labeling panel
Label as Spambot
Label as Genuine
Exit
t-SNE2
t-SNE1
Accounts
Tweets
WordCloud
Manoj Cheenath
@cheenath
Joined April 27, 2008
Tweets: 329
Following: 445
@yifany
Joined June 11, 2008
Tweets: 2419
Following: 142
Steven Menzel
@smenzel5
Joined April 5, 2009
Tweets: 2864
Following: 925
anita wray white
@dylangirl99
Joined September 16, 2009
Tweets: 2002
Following: 2002
Dass
@dass90xz
Joined December 20, 2009
Tweets: 2237
Following: 595
Polarity
TopicID
tweet_count
retweet_count
reply_count

12

Upcoming NSF-funded project: Data Fusion for Sea Ice Classification

- SAR imagery
- Sentinel-1
- NISAR
- IceBridge
- ICESat
- ICESat-2



NSF

University of Colorado Boulder

University of Colorado Denver

NSIDC National Snow and Ice Data Center

EarthCube Data Capabilities: Enabling Analysis of Heterogeneous, Multi-source Cryospheric Data

- Morteza Karimzadeh, Geography, Information Science (CU Boulder)
- Farnoush Kashani-Banaei, Computer Science (CU Denver)
- Andrew Barrett (NSIDC)
- Walt Meier (NSIDC)
- Siri Jodha Khalsa (NSIDC)

NSF

University of Colorado
Boulder

University of Colorado
Denver

NSIDC
National Snow and Ice Data Center

Thank you!

Q/A

Karimzadeh@colorado.edu

@mortezakz

IceNet: A seasonal, deep learning-based pan-Arctic sea ice forecasting system

Tom Andersson

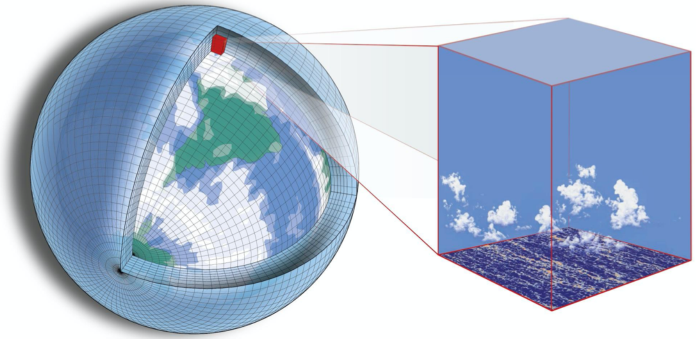
Scott Hosking, María Pérez-Ortiz, Brooks Paige, Chris Russell, Andrew Elliott , Stephen Law, Tony Phillips, Jeremy Wilkinson, Yevgeny Askenov, Bablu Sinha, Will Tebbutt, Fruzsina Agocs, and Emily Shuckburgh

British Antarctic Survey, Alan Turing Institute, Cambridge University, UCL Centre for AI, National Oceanography Centre

Two climate forecasting paradigms: Physics-driven vs. data-driven

<u>Dynamical models (physics-driven)</u>

- Model the laws of physics directly
- Based on causality
- Computationally expensive

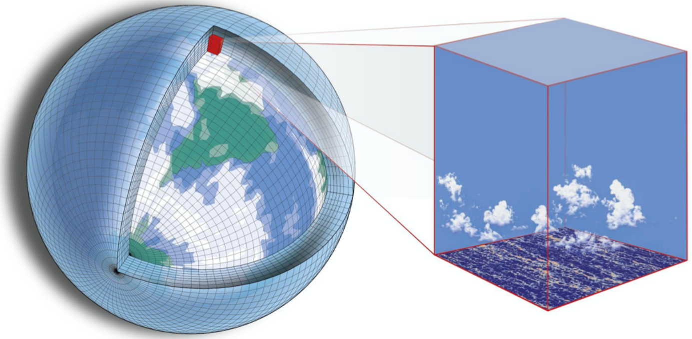


Credit: Schneider et al., Nature Climate Change

Two climate forecasting paradigms: Physics-driven vs. data-driven

<u>Dynamical models (physics-driven)</u>

- Model the laws of physics directly
- Based on causality
- Computationally expensive



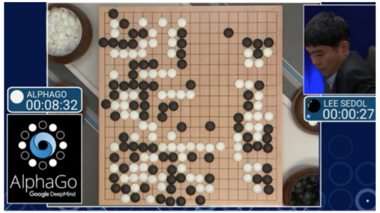
Credit: Schneider et al., Nature Climate Change

<u>Statistical models (data-driven)</u>

- Automatically learn complex, non-linear relationships between variables from raw data
- Based on correlations
- Computationally cheap (once trained)

Credit: Shutterstock

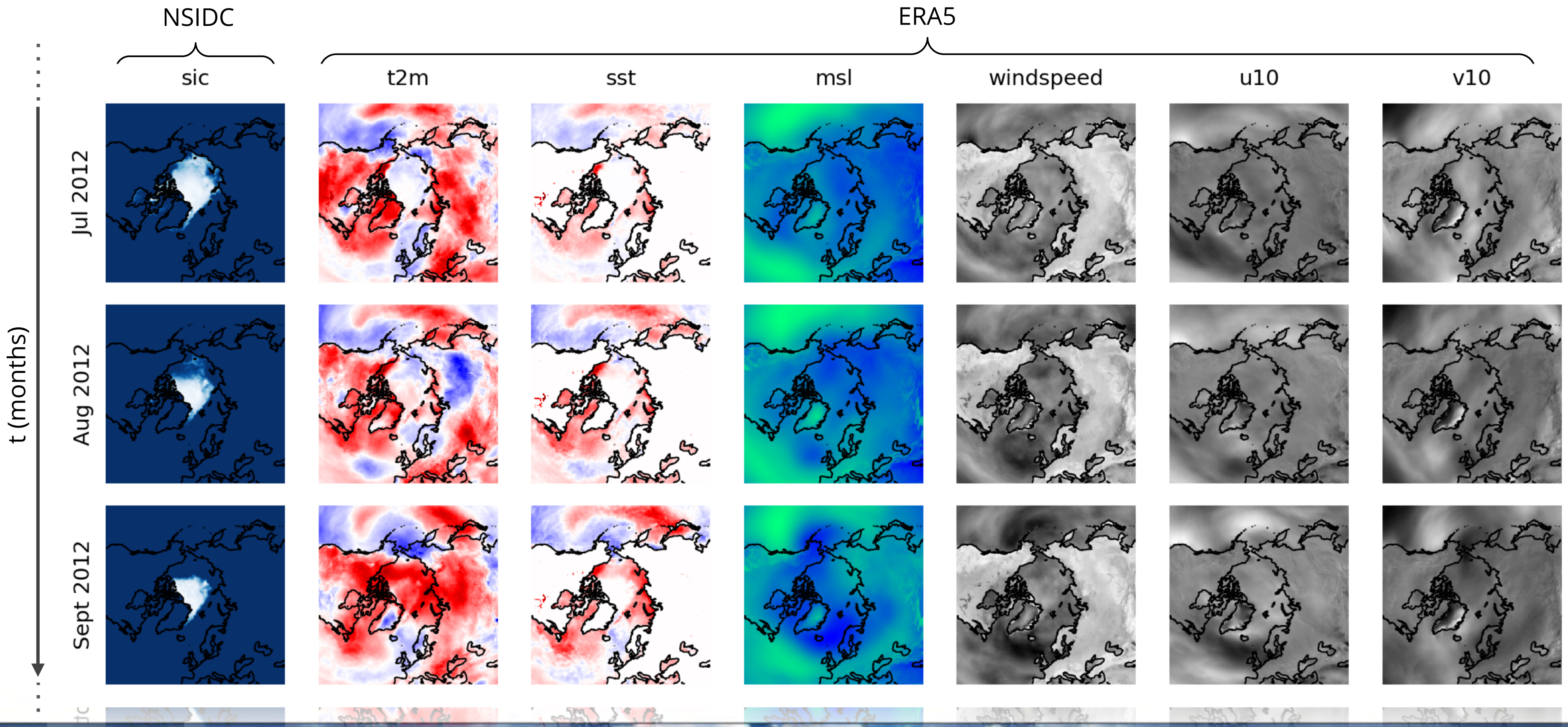
Credit: Vinyals et al., CVPR



Credit: DeepMind

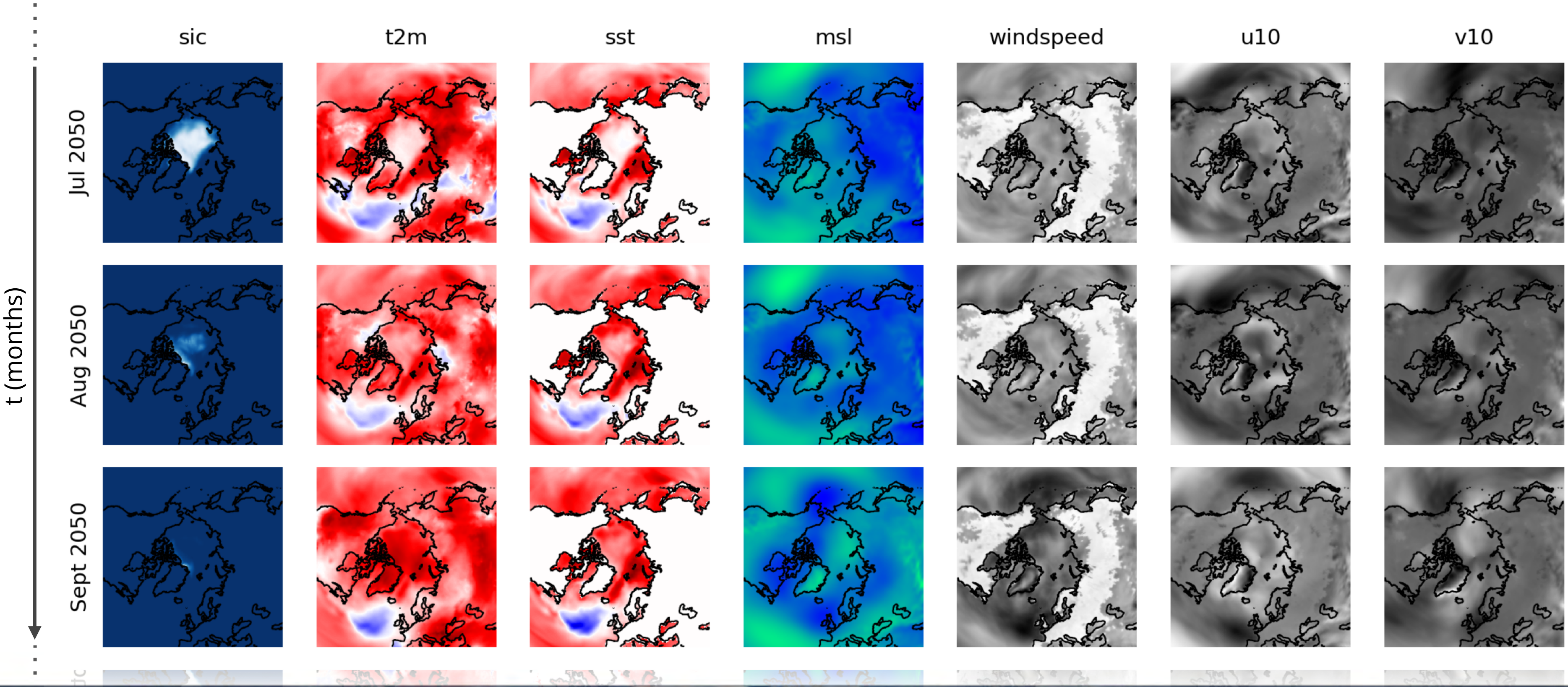
IceNet data: Observations

Time period: 1979-present (500 months)

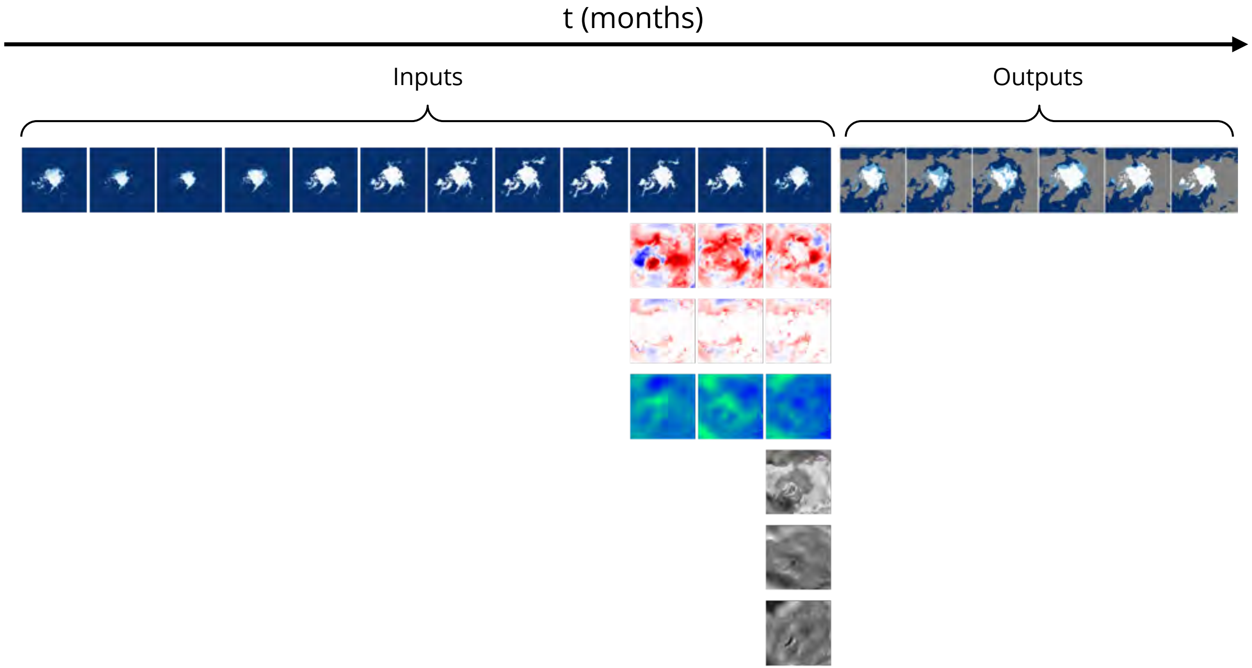


IceNet data: Climate model (MRI-ESM2.0)

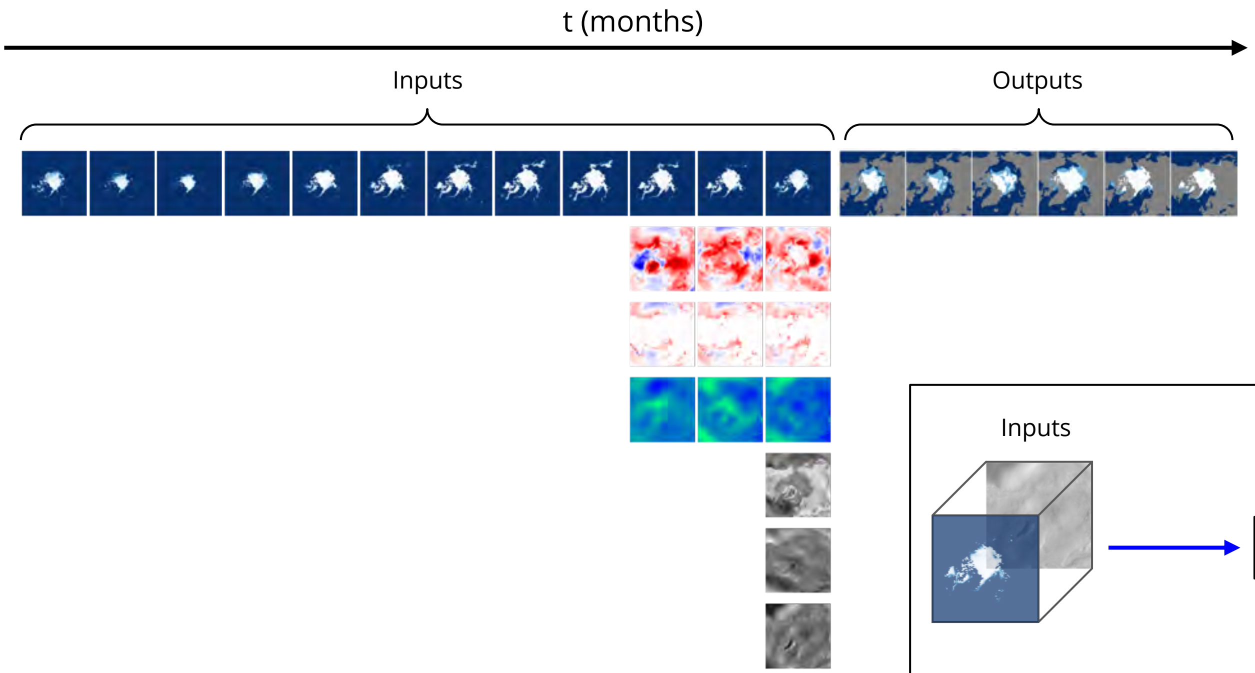
Time period: 1850-2100 (3012 months)

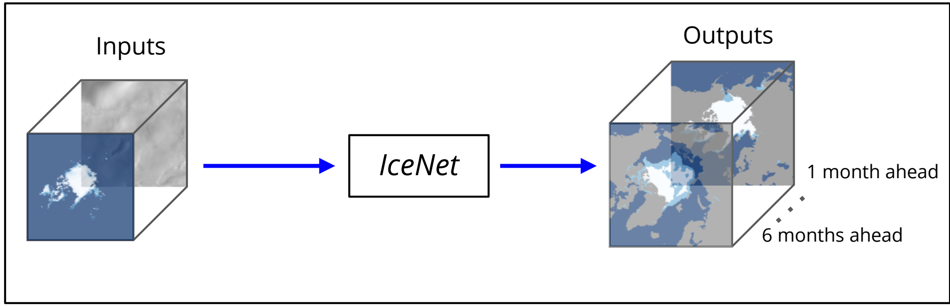


IceNet design: Inputs and outputs

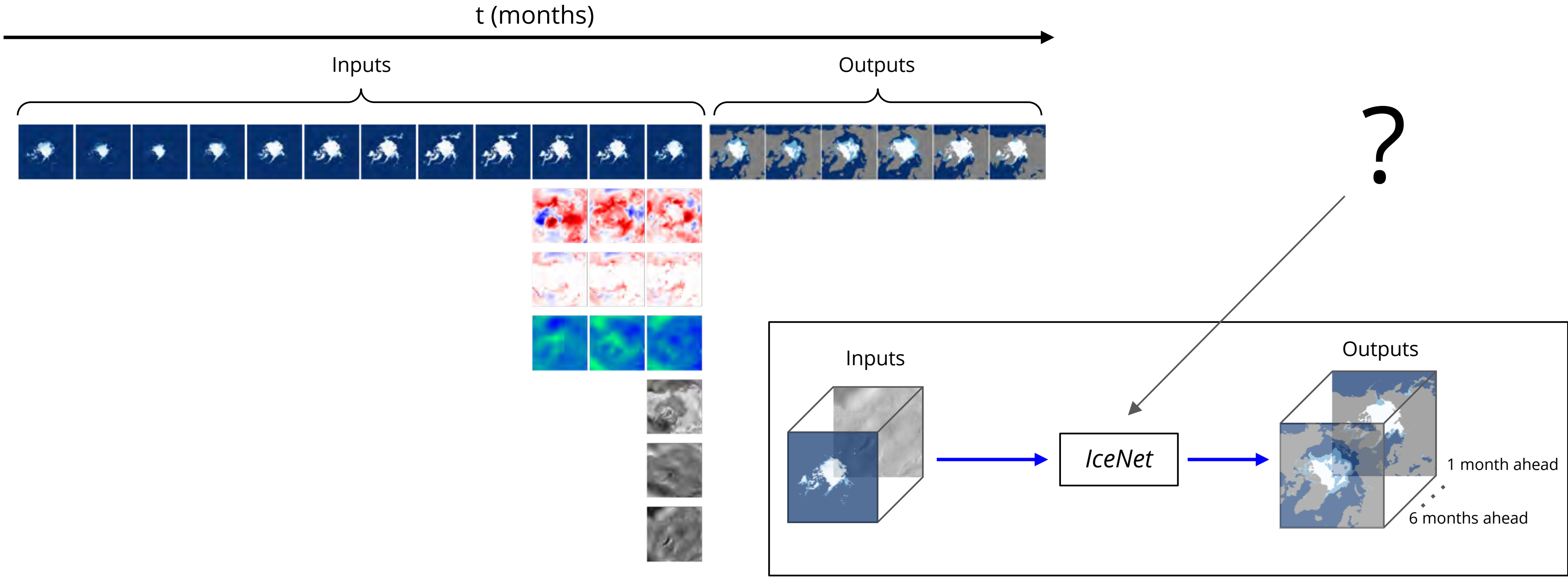


IceNet design: Inputs and outputs



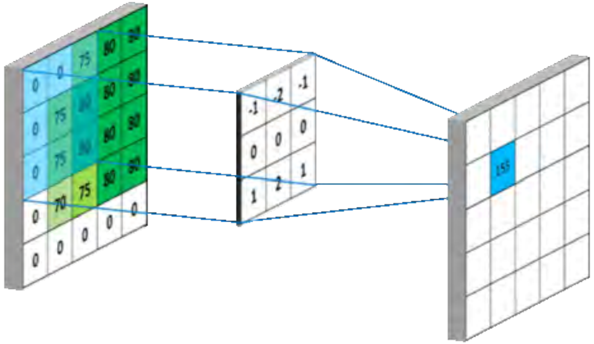


IceNet design: Inputs and outputs

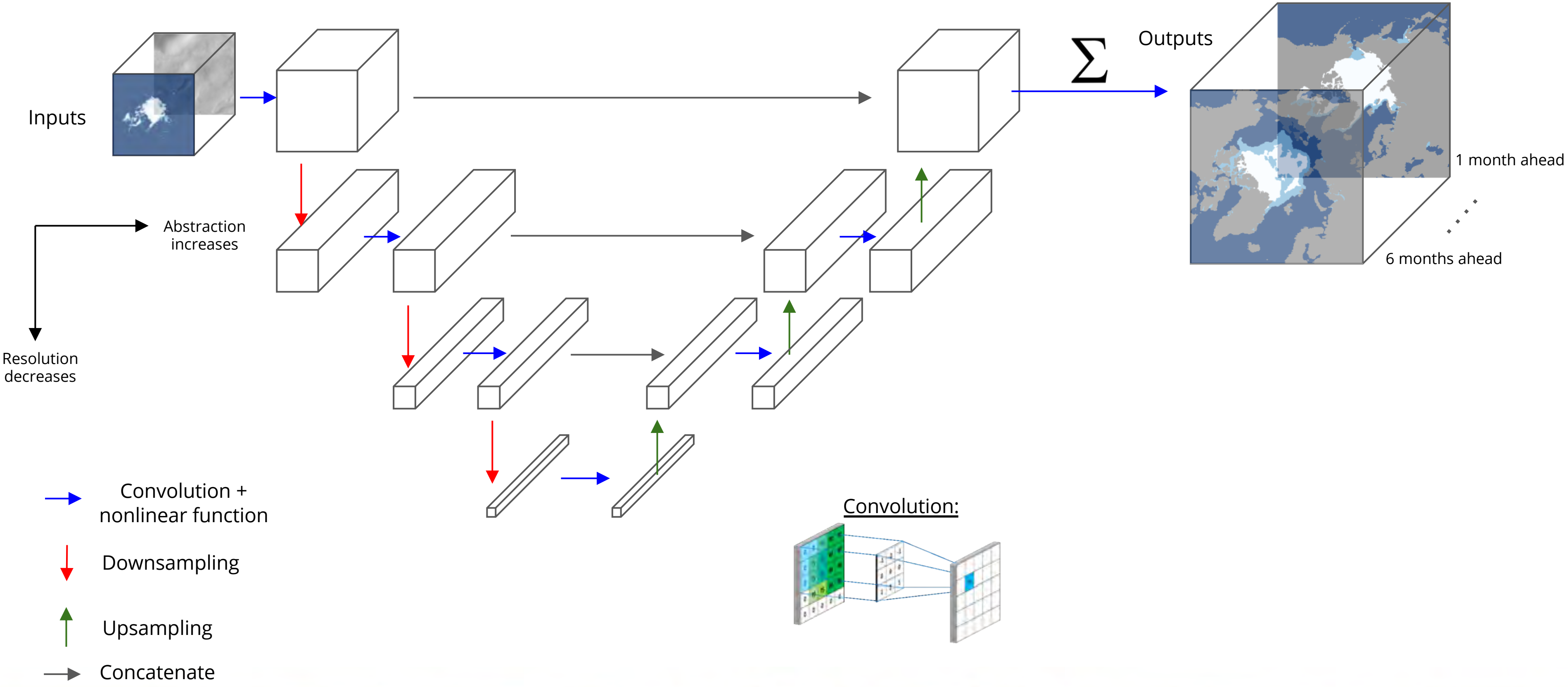


IceNet design: U-Net Architecture

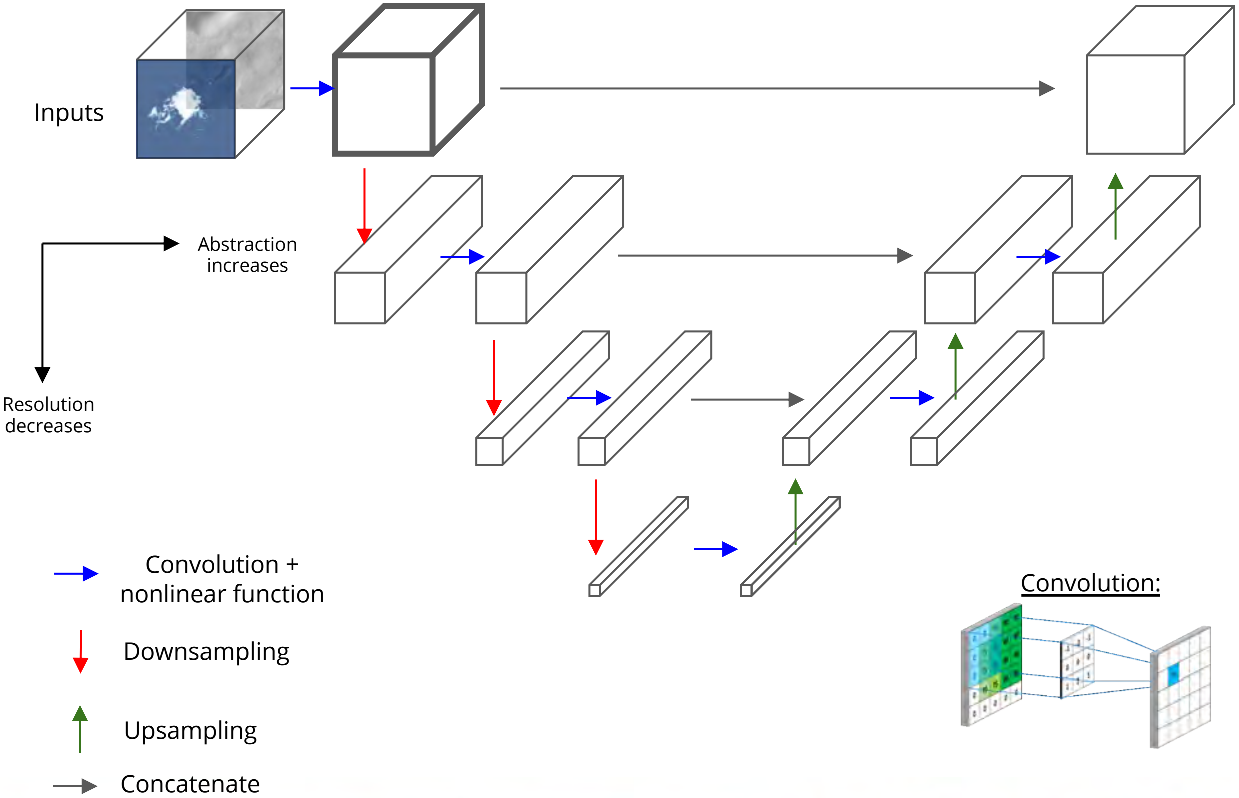
<u>2D Convolution:</u>



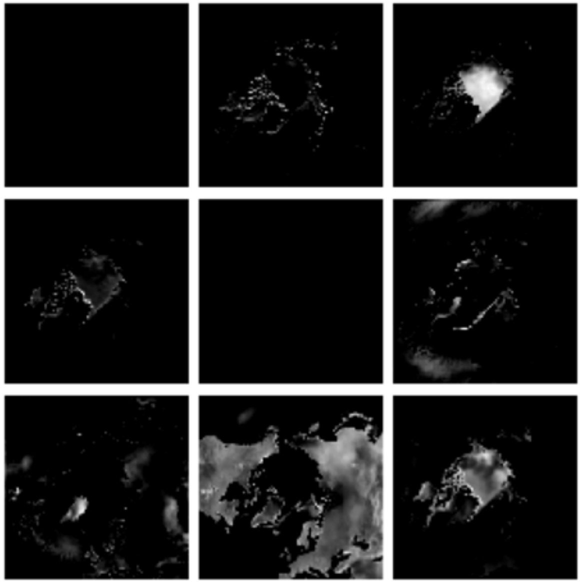
IceNet design: U-Net Architecture



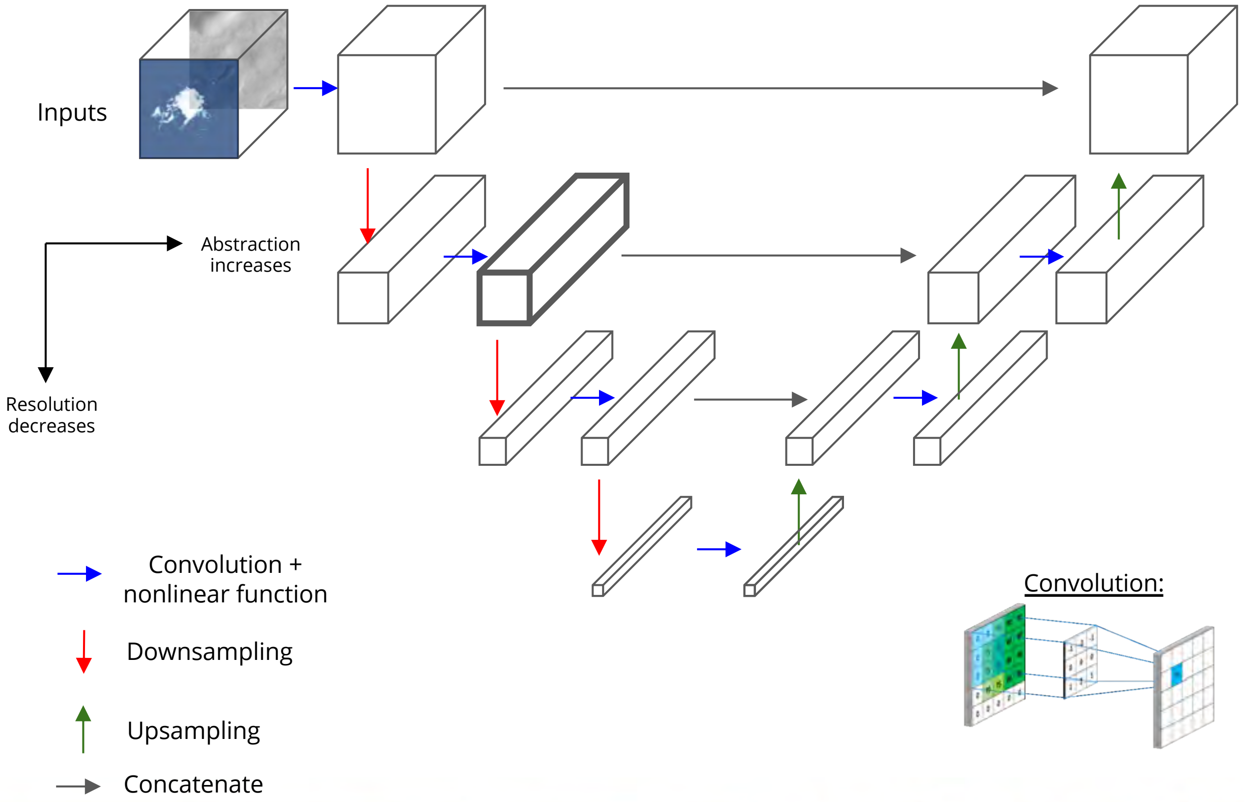
IceNet design: U-Net Architecture



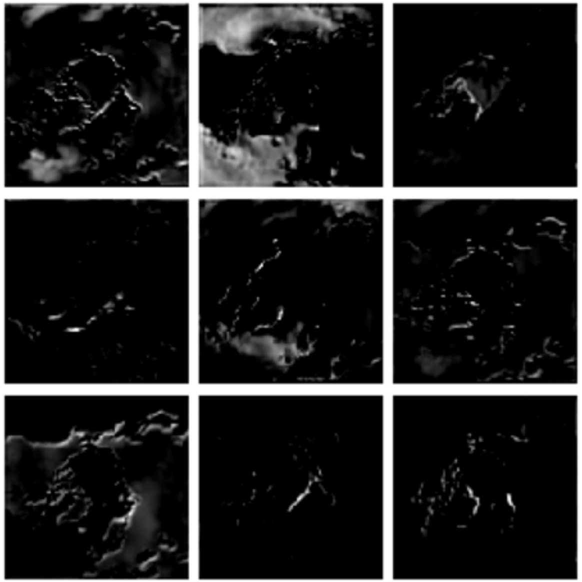
Activation maps



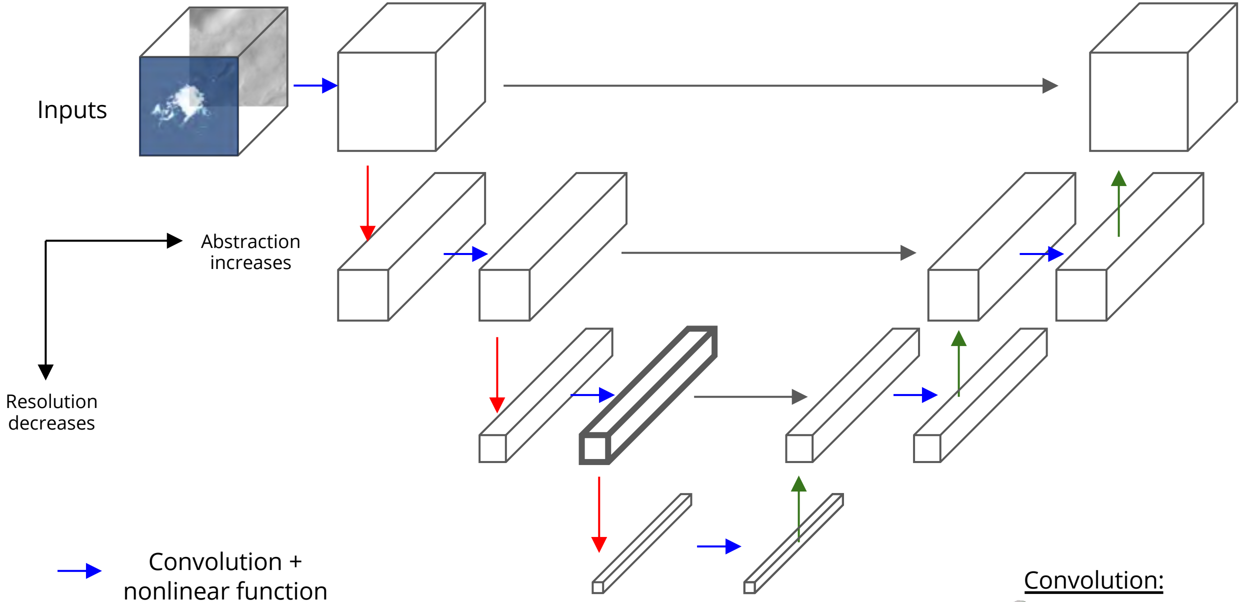
IceNet design: U-Net Architecture



Activation maps

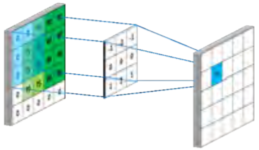


IceNet design: U-Net Architecture

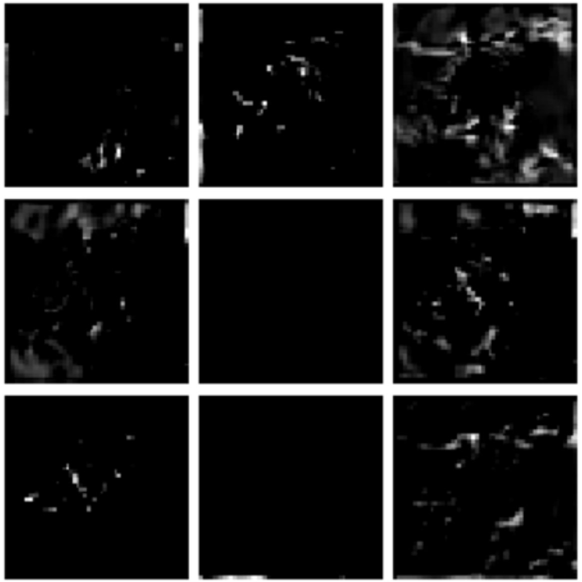


- Convolution + nonlinear function
- Downsampling
- Upsampling
- Concatenate

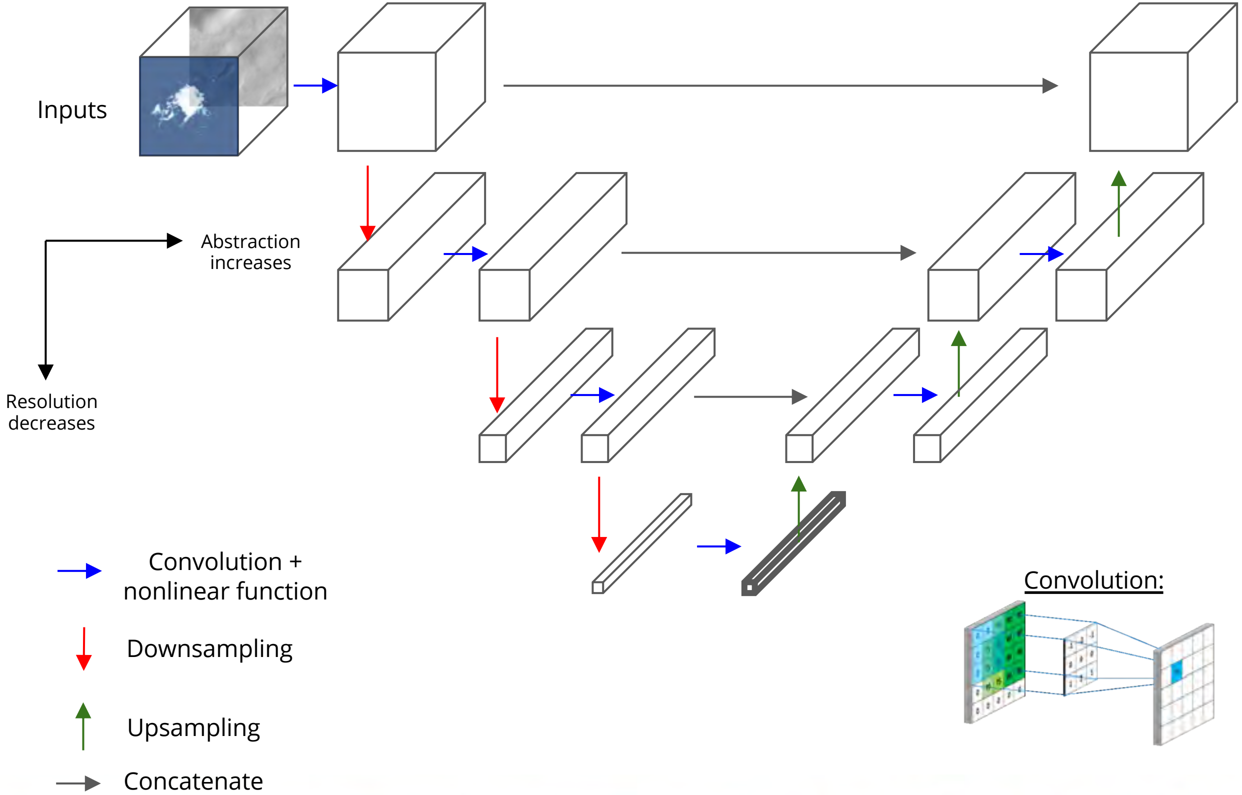
Convolution:

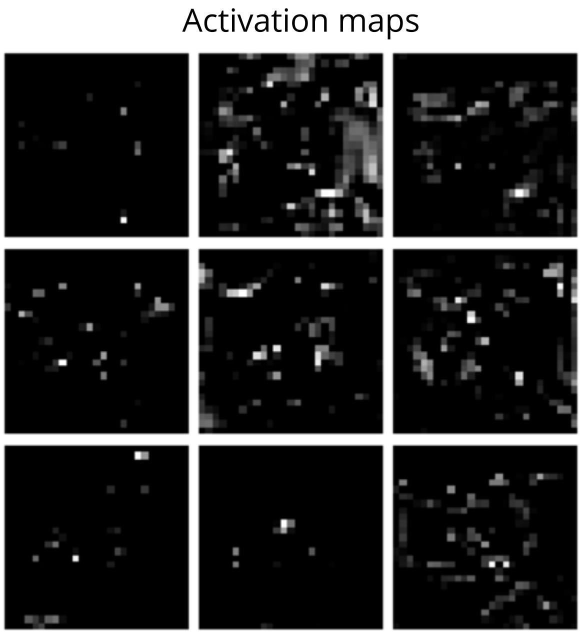


Activation maps

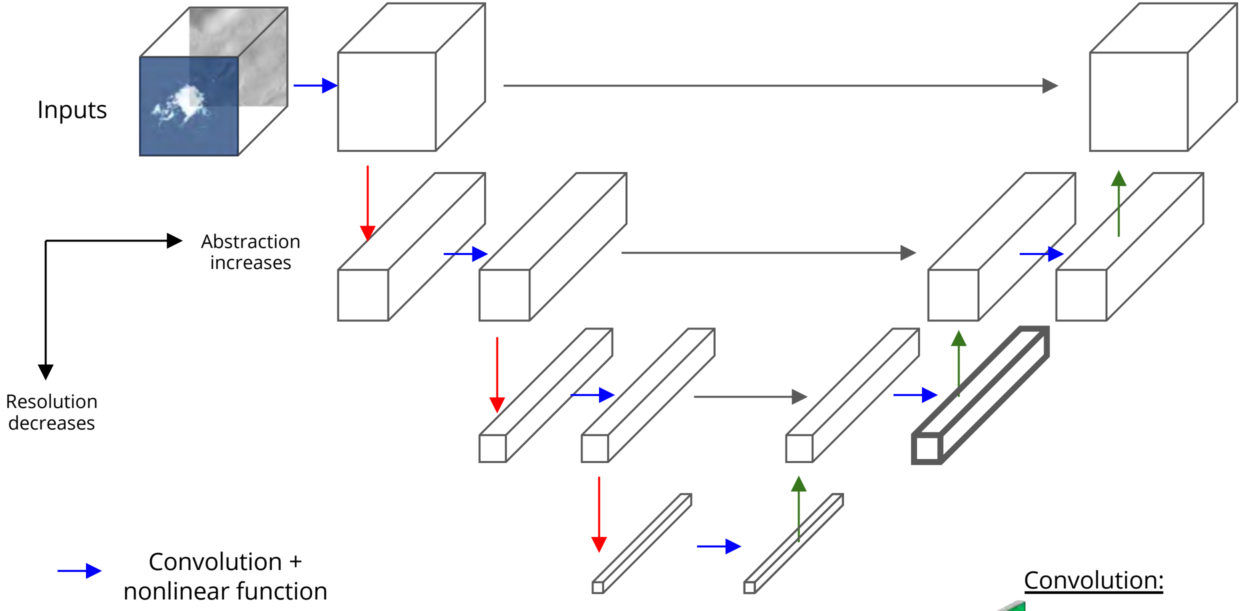


IceNet design: U-Net Architecture





IceNet design: U-Net Architecture



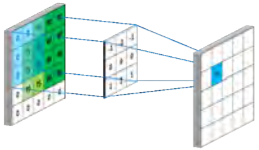
Convolution + nonlinear function

Downsampling

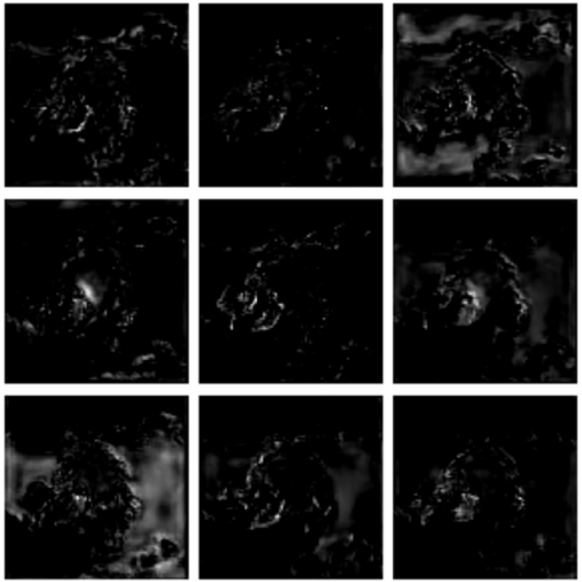
Upsampling

Concatenate

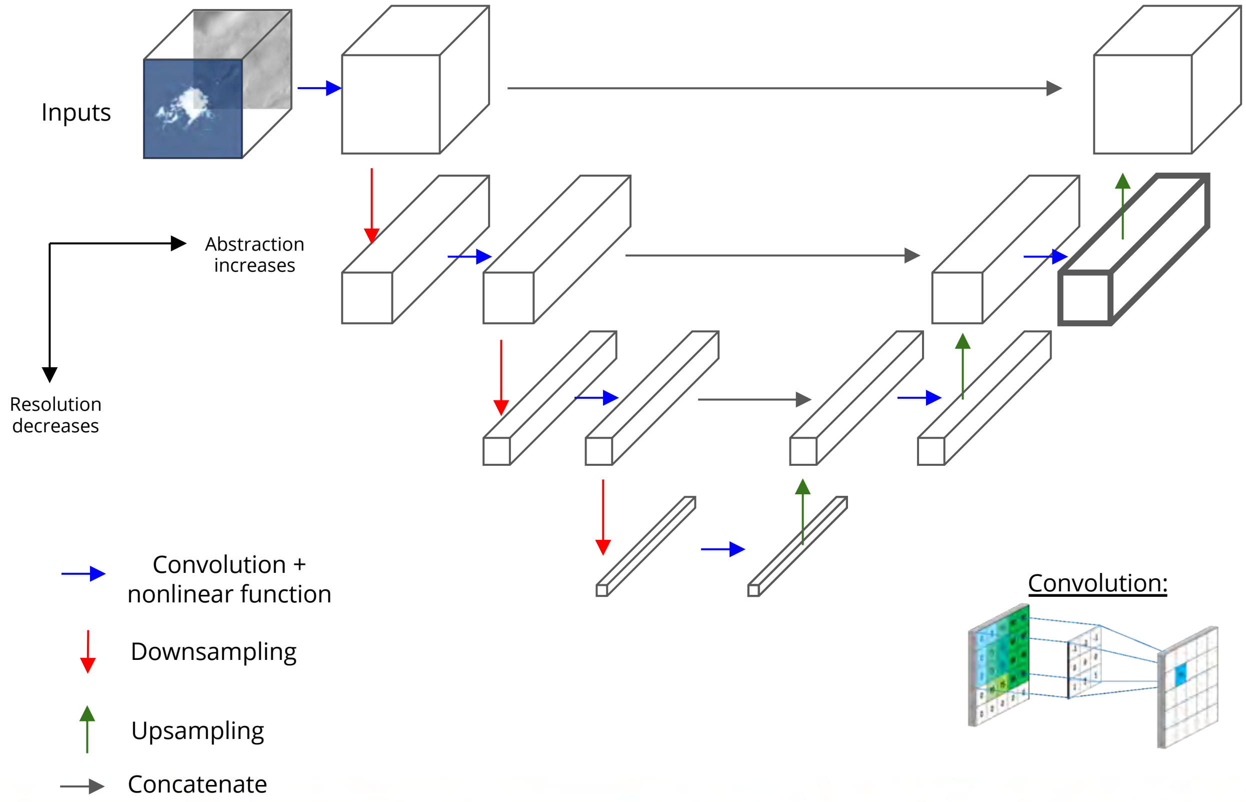
Convolution:



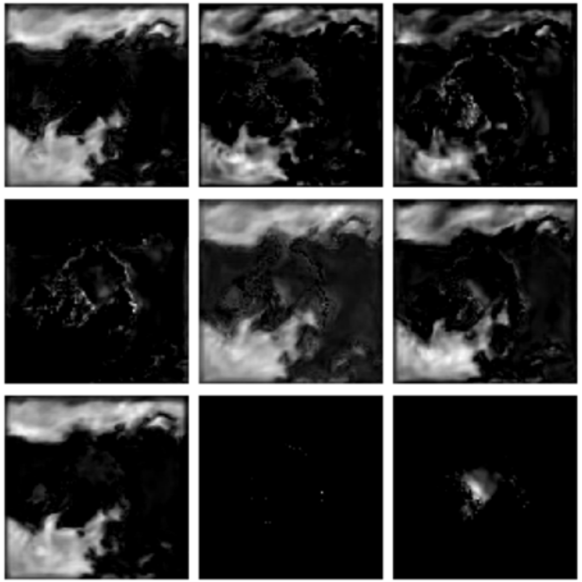
Activation maps



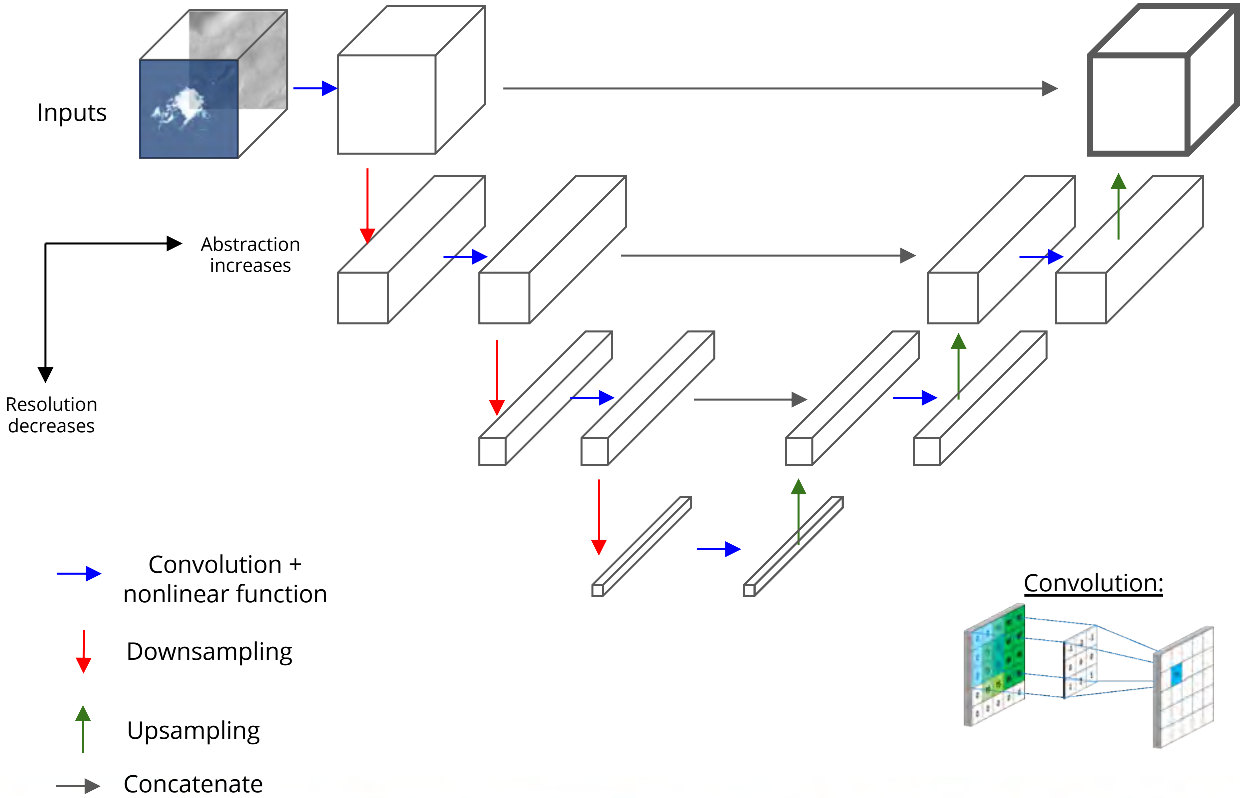
IceNet design: U-Net Architecture



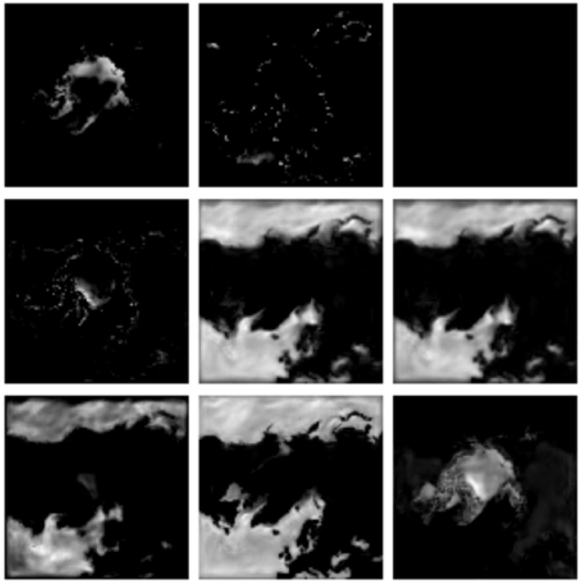
Activation maps



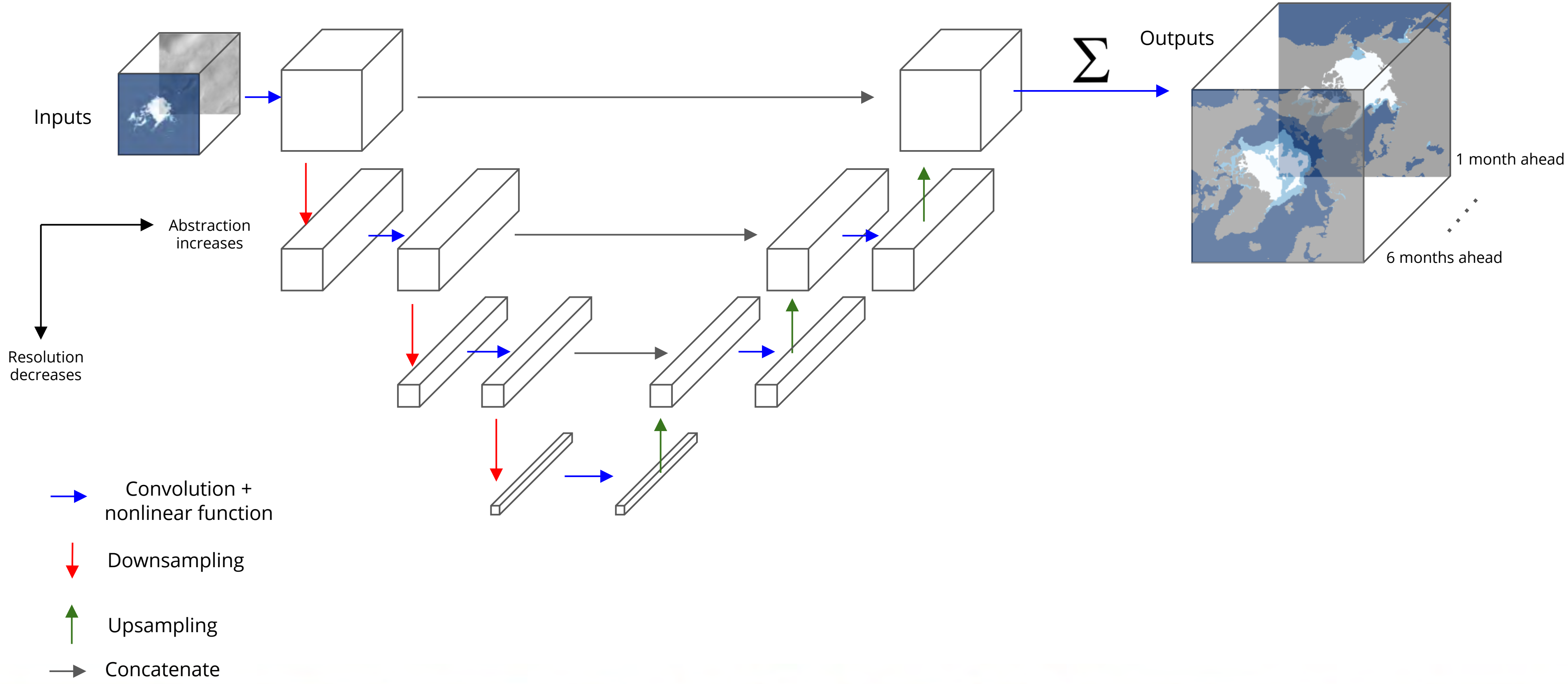
IceNet design: U-Net Architecture



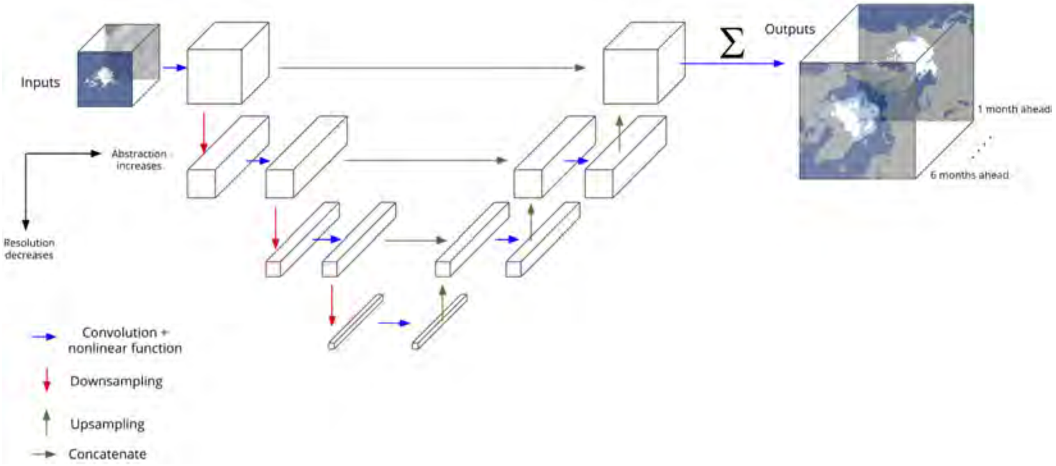
Activation maps



IceNet design: U-Net Architecture



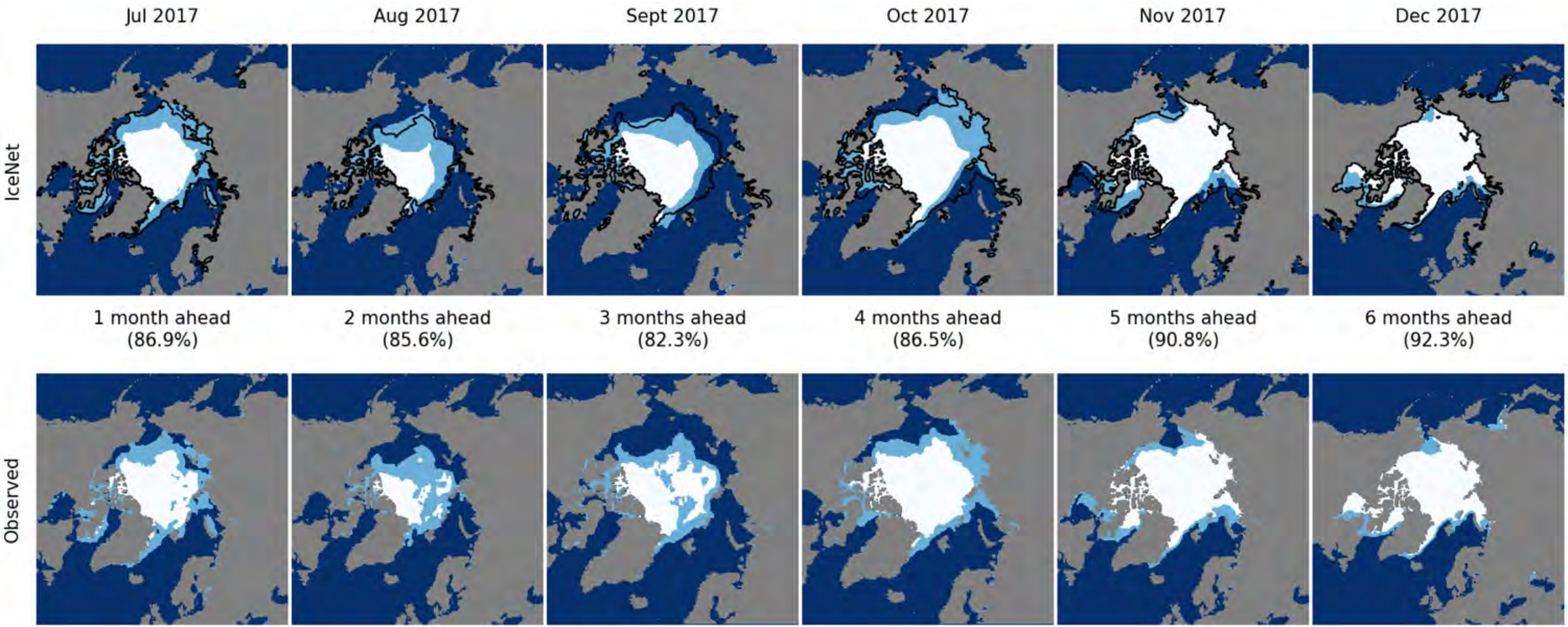
IceNet design: U-Net Architecture



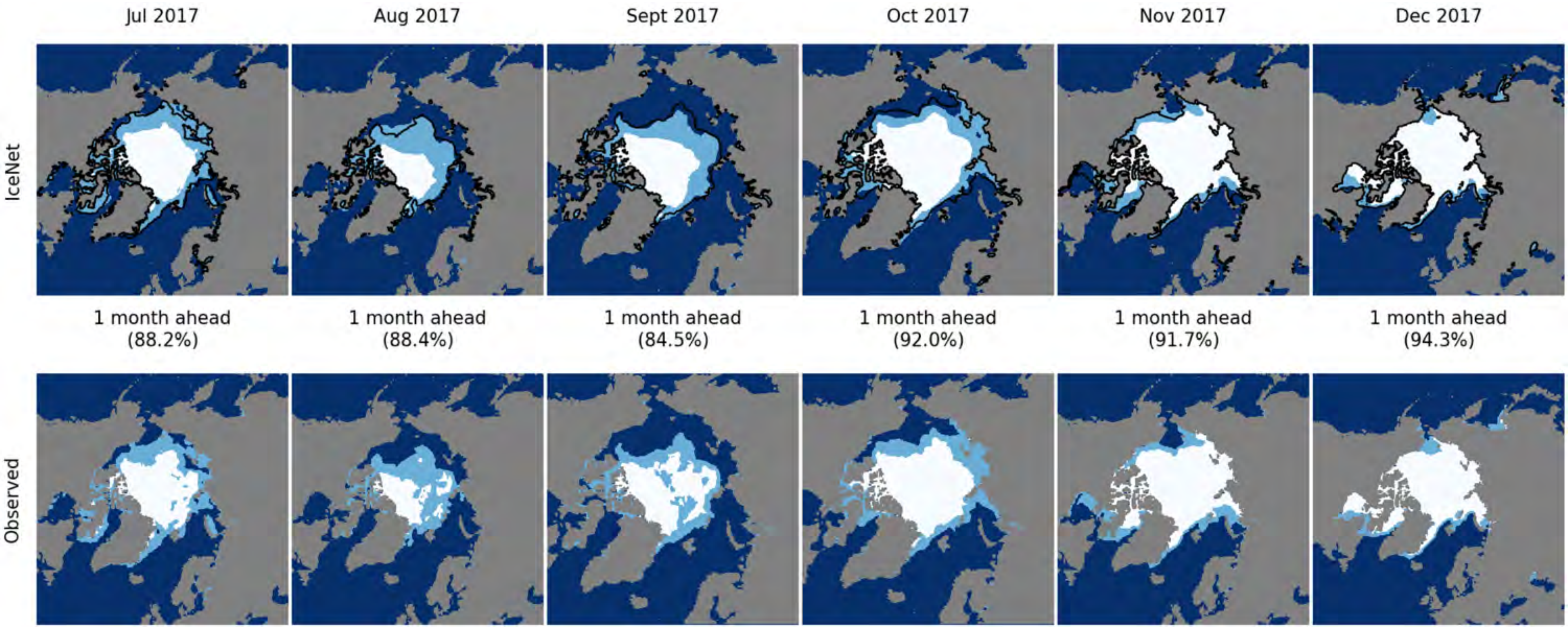
- Three output classes:
  a. No ice (SIC < 15%)
  b. Marginal ice (15% < SIC < 80%)
  c. Full ice (SIC > 80%)

- # of params: 10, 983, 434
- Pre-train on >10,000 months of climate model data (MRI-ESM2.0)
- Fine-tune on 1979-2015 observational data
- Validate (hindcast) on 2016-2018
- Ensemble of 3 networks

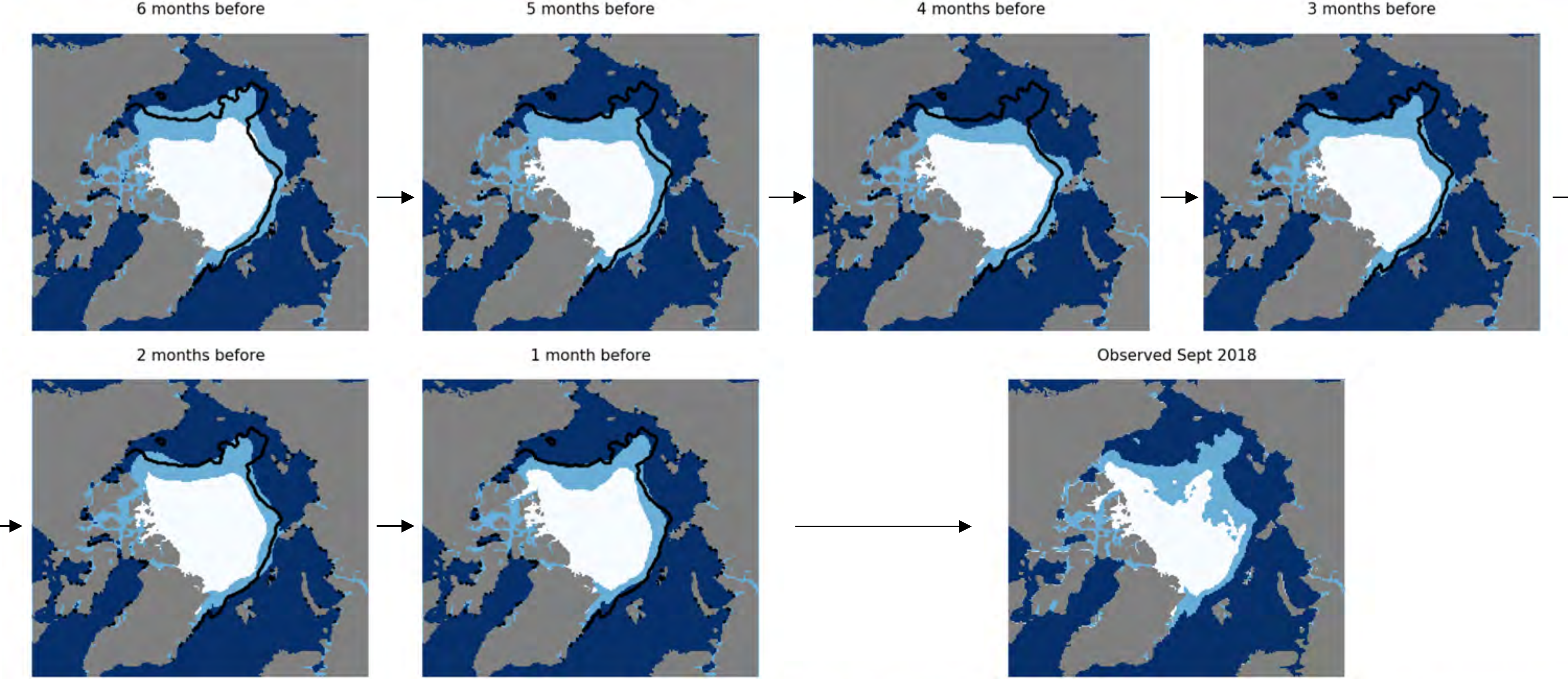
IceNet predictions: Predict entire second half of 2017 starting in June



IceNet predictions: Predict second half of 2017 one month ahead



IceNet predictions: September 2018

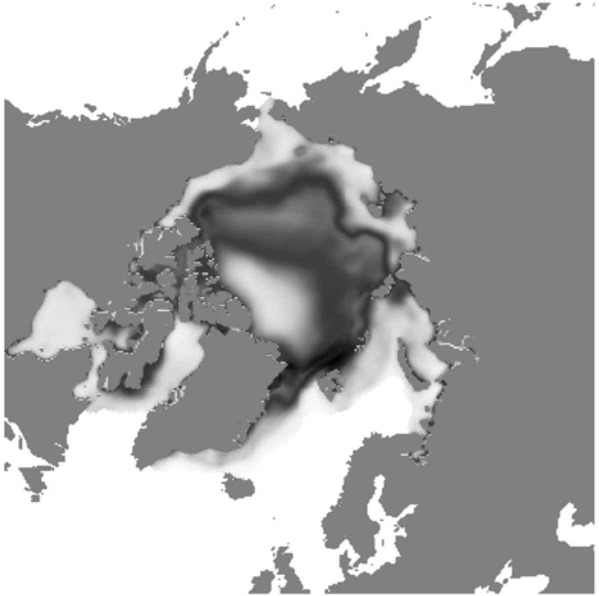


IceNet predictions: Prediction uncertainty (Aug 2017)

p(ice) = p(marginal ice) + p(full ice)

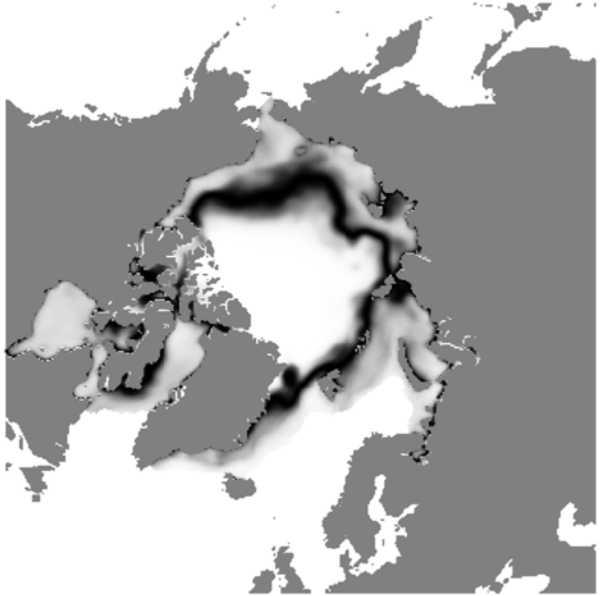
Observed

3 class entropy



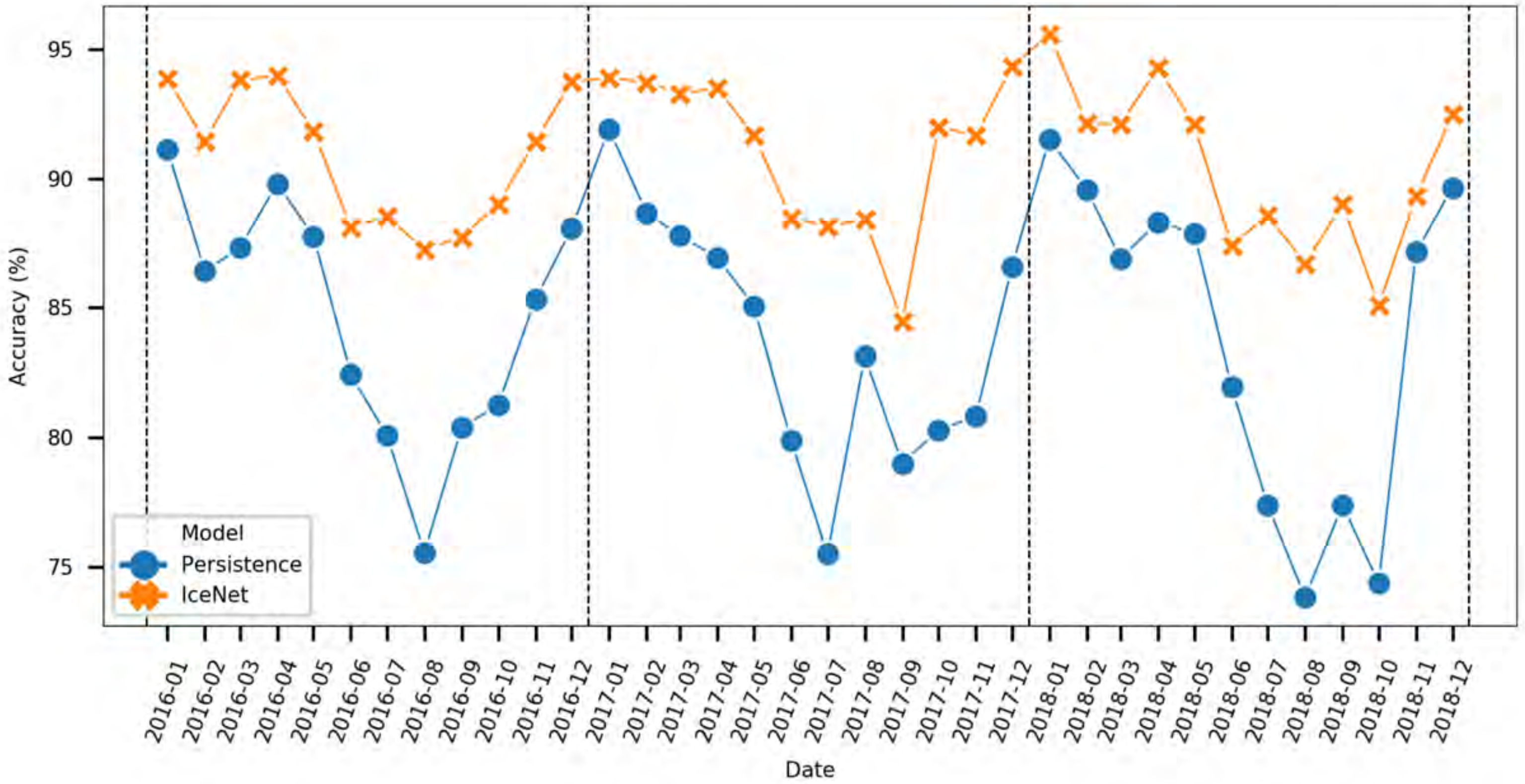
1 month ahead

2 class entropy

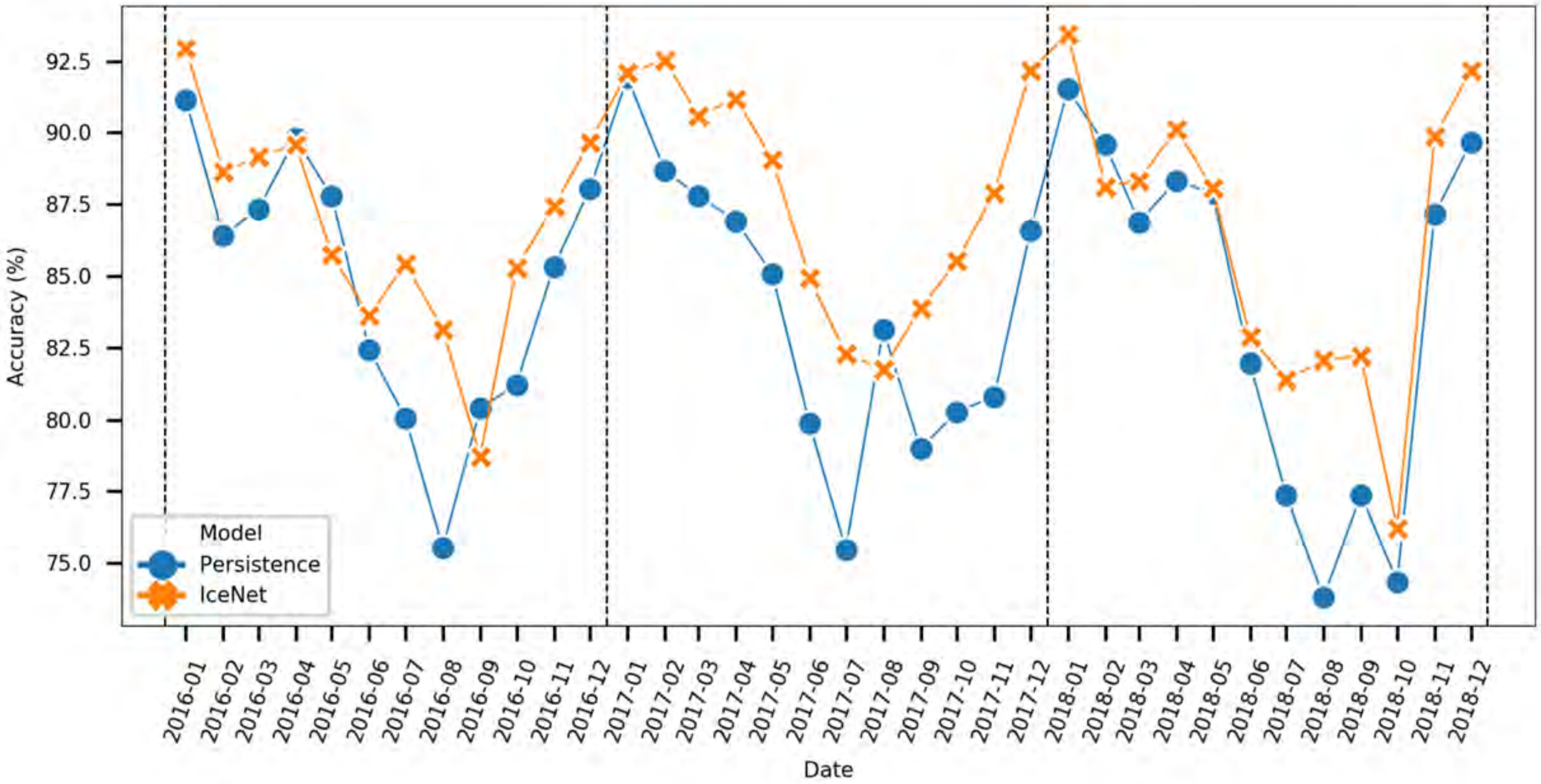


1 month ahead

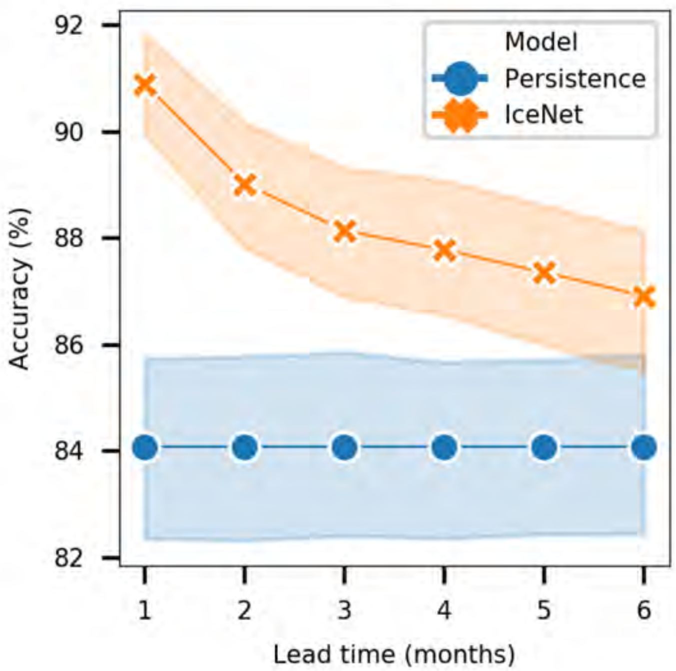
Hindcast results: 1 month ahead

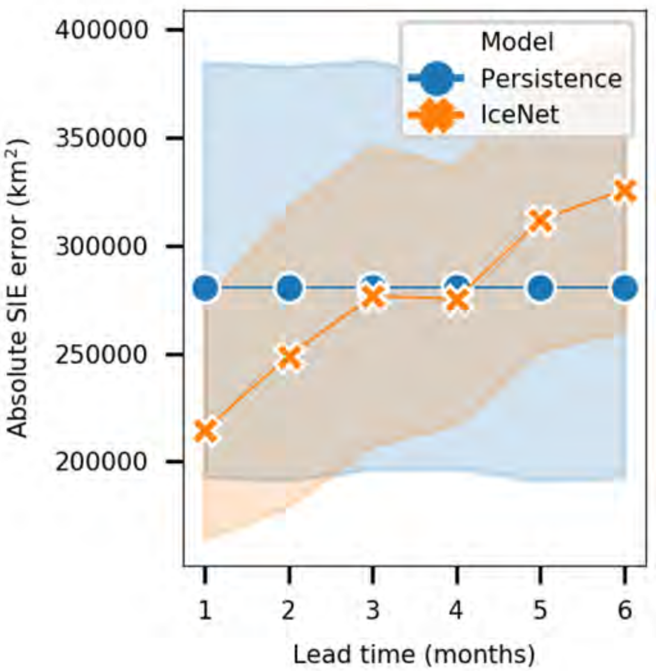


Hindcast results: 6 months ahead



Validation mean performance vs. lead time





Thanks for listening!

Contact: tomand@bas.ac.uk

Entropy